Predicting Mathematics Achievement: A Machine Learning Approach Using TIMSS 2015 U.S. National Public-Use Data

Word count: 1995

Abstract

Using TIMSS 2015 U.S. national public-use data, this study shows how machine learning could be applied to large-scale international studies. The analysis provides a comparison of six commonly used machine learning models in identifying students with low mathematics performance, which would allow educators to effectively allocate resources to intervene and thus improve their performance. The results show that logistic regression, elastic net, and extreme gradient boosting all perform well in terms of balanced accuracy and sensitivity. Leveraging the high dimensionality of the data, this research analyzes 142 student and school variables and identifies characteristics that are associated with mathematics performance, including students' home education resources, attitudes toward mathematics, and use and possession of computers or tablets.

Keywords

Machine learning; TIMSS; Mathematics achievement; Large-scale educational research

Introduction

Objective

Poor mathematics achievement creates difficulties for students in everyday activities and leads to serious consequences in educational attainment and career advancement (Jordan, 2010; Geary, 2011). Identifying characteristics that could predict poor mathematics achievement would help educators allocate resources effectively and intervene in time. Using 2015 U.S. data from the Trends in International Mathematics and Science Study (TIMSS), this paper attempts to 1) provide a comprehensive evaluation of machine learning (ML) methods in analyzing international large-scale education data, and 2) identify student and school background characteristics that have the largest importance in the prediction.

Theoretical Framework

Existing literature has explored various factors that may be associated with students' mathematics achievement. Characteristics ranging from students' gender (McGraw et al., 2006), race (Brown-Jeffy, 2009), and family background (Wang, 2004) to teacher qualification (Darling-Hammond, 2000) and school resources (Burtless, 2011), to name a few, are believed to be significant determinants of achievement. However, most studies have employed conventional statistical methods and have focused on only a few theory-driven variables.

With the recent development of machine learning (ML), a number of data-driven techniques are now utilized by education researchers in areas including predicting dropout (Sansone, 2017) and course score (Polyzou & Karypis, 2016). However, only a handful of studies have focused on applying ML techniques to large-scale international assessment data (Gabriel & Jason Signolet, 2018; Yoo, 2018), and none have compared the performance of different ML methods.

Significance

This study thus attempts to fill the research gaps identified above by comparing the performance of the six most common ML algorithms using TIMSS 2015 U.S. national public-use data. The methodology presented could also be applied to the analysis of data from other education systems. The study also contributes to the literature by leveraging the high dimensionality of the data, as it investigates as many student and school variables as possible to identify characteristics associated with mathematics performance. Being able to accurately predict students who would have low mathematics performances would allow in-time intervention to increase performance.

Data Source

TIMSS is an international comparative study designed to measure trends in mathematics and science achievement at the 4th and 8th grades, as well as collect information about educational contexts that may be related to student achievement (Provasnik et al., 2016). In 2015, 64 education systems participated in TIMSS. This paper analyzes data from the United States, specifically, the 8th-grade student achievement data as well as student and school background questionnaire data from the U.S. national public-use data file. The national public-use data file contains additional U.S.-specific information, such as students' race/ethnicity, which provides more context for data analysis.

Methods

Data preprocessing

As Figure 1 shows, after merging the student achievement data with student and school background questionnaire data, there were 10,221 observations with 572 variables. A binary outcome variable was created that equals 1 if the student is considered low-performing in mathematics (defined as having the average of all 5 plausible values for mathematics below the intermediate international benchmark) and 0 otherwise.[1]

The data cleaning process included removing irrelevant variables (e.g., ID-related variables, file maintenance variables, weights, and plausible values), removing variables that have more than 15% missing data, removing variables that are highly correlated with each other, and standardizing variables. A more detailed data cleaning process is documented in Appendix A.

The data cleaning process resulted in 5,516 grade 8 students (54% of the original sample) with 143 student and school variables, as documented in Appendix B. As Table 1 shows, using the outcome variable as the stratifier, the remaining observations were randomly split into training and test datasets with the conventional ratio of 8:2.

---

[1] TIMSS defines four levels of student achievement, referred to as international benchmarks: Advanced, High, Intermediate, and Low. These international benchmarks provide a way to understand how students' proficiency in mathematics varies at different points on the TIMSS scale. Grade 8 students who perform below 400 are considered below the low international mathematics benchmark. Grade 8 students who perform at or above 400 but below 475 are considered to be at the low international mathematics benchmark (Mullis, Martin, Foy, & Hooper, 2016).

The training set was used for model construction and validation, where 5-fold cross-validation[2] is performed to tune hyperparameters that make the model perform the best on the training set. We then applied the tuned models to the test set for model evaluation, using a variety of the performance metrics detailed in the Results and Discussion section.

Machine Learning Models: Brief Introduction

Machine learning is the science of giving computers the ability to learn without being explicitly programmed (Samuel, 1959). This paper implemented six common ML algorithms to approach a supervised problem.[3] The following section briefly describes each of these algorithms.[4]

Logistic Regression

Logistic regression is one of the most commonly used ML algorithms for binary classification. A sigmoid function is applied to map the linear regression output to continuous values from 0 to 1, which represent the probability of classification. A threshold classifier of 0.5 is used to convert the probability into either 0 or 1.[5]

Elastic Net

Elastic net is an extension of logistic regression. When the number of predictors becomes large, overfitting and multicollinearity become problems, which often plague logistic regression with degeneracies (Hastie & Qian, 2014). Penalized regression methods, such as elastic net, alleviate the problem by adding a penalty term to the model,

$$\lambda \sum_{j=1}^{P} (\alpha|\beta_j| + (1-\alpha)\beta_j^2)$$

where λ controls the overall strength of the penalty and α controls the elastic net penalty.[6]

Decision Tree

The decision tree model is one of the tree-based methods, which involve splitting the dataset recursively.[7] At each step, the split is made based on the independent variable that results in the largest possible reduction in heterogeneity of the dependent variable.[8]

Random Forest

Random forest is an extension of the decision tree model. It involves producing and combining multiple trees (hence "forest") to yield a single consensus prediction.[9]

---

[2] Technical details on cross-validation are provided by Hastie et al. (2001).

[3] Machine learning algorithms can be divided into two main categories: supervised learning, where there is a defined output variable; and unsupervised learning, where the model is designed to find patterns.

[4] A more detailed technical explanation is provided by James et al. (2014).

[5] The logistic regression model is implemented by the "glm" R package (R Core Team, 2013).

[6] When $\alpha = 1$, it is the LASSO (Least Absolute Shrinkage and Selection Operator) model, and when $\alpha = 0$, it is the Ridge model. The elastic net model is implemented by the "glmnet" R package. (Friedman, Hastie, & Tibshirani, 2010).

[7] Splitting a dataset recursively means that the subsets that arise from a split are further split until a predetermined termination criterion is reached.

[8] The decision tree model is implemented by the "rpart" R package (Therneau & Atkinson, 2018).

[9] It is known that combining a large number of trees can often result in dramatic improvements in prediction accuracy, at the expense of some loss in interpretation. The random forest model is implemented by the "randomForest" R package (Liaw & Wiener, 2002).

Extreme Gradient Boosting

Extreme gradient boosting, like random forest, uses an ensemble of multiple trees to create a more powerful prediction model. Unlike random forest, which builds and combines a forest of randomly different trees in parallel, extreme gradient boosting builds a series of trees in such a way that when each tree is trained, it attempts to correct the mistakes, or misclassification, of the previous tree in the series. The final prediction is based on the aggregate results of all trees in the series. The model also reduces the risk of overfitting by heavily penalizing complexity.[10]

Neural Network

Neural network models are inspired by the way biological neural networks in the human brain process information. A simple form of neural network is called multi-layer perceptron, which has three components: an input layer, one or many hidden layer(s), and an output layer.[11]

Results and Discussion

Research Question 1: How do the selected machine learning methods compare in predicting low mathematic performers?

While accuracy is usually used as a performance metric, it would be biased when the data set is unbalanced (Zahedifard et al., 2015), as in this case. For example, a baseline model, which uses the simplest rule in predicting outcomes by always predicting the most frequent category, would predict that all 1,103 observations in the test set are in the "non-low-mathematics achiever" category and none are in the "low-mathematics achiever" category. This would have an accuracy of 852/1,103 = 0.772. In comparison, balanced accuracy, which is the harmonic mean of precision and recall, has proven to be useful in evaluating models with unbalanced categories (García, Mollineda, & Sánchez, 2009). Therefore, we primarily use balanced accuracy to evaluate and compare the models. The same baseline model would have a balanced accuracy of ½ * (852/852 + 0/251) = 0.5. Table 2 lists additional metrics for finer details, including sensitivity/recall, specificity, positive predicted value/precision, and negative predicted value.

All six ML models performed above the baseline model. The decision tree model yielded a balanced accuracy of 65.5%, the lowest across all six models. Among the tree-based models, the random forest and extreme gradient boosting models showed improved performance against the decision tree model, with balanced accuracy of 70.0% and 74.9%, respectively. Logistic regression yielded the highest balanced accuracy (76.0%), 10 percentage points higher than the decision tree model and 26 percentage points higher than the baseline model.

Among the other metrics reported, sensitivity measures the percentage of low-performing grade 8 students who are correctly identified as being at risk of performing poorly in mathematics. And specificity measures the percentage of non-low-math-achiever students who are correctly identified as not at risk of performing poorly in mathematics. Given the context, it should be of policy interest for schools and other stakeholders to expect higher sensitivity from the model. In other words, it would be ideal to have as few cases as possible where students are predicted to be non-low-mathematics achievers when in fact they are at high risk of performing poorly. Logistic regression, again, topped all selected models in sensitivity (58.2%).

[10] The extreme gradient boosting model is implemented by the "xgboost" R package (Chen, et al., 2018).
[11] A single hidden layer neural network model is implemented by the "nnet" R package (Venables & Ripley, 2002).

Research Question 2: Which variables have the most predictive value in predicting low mathematics performance?

Besides returning balanced accuracy for the test data, each of the six models computes variable importance scores for 142 predictor variables. Although the technical details of calculating the variable importance estimates differ across models (Kuhn, 2017), all methods, in essence, rank predictor variables based on the role each plays in predicting the outcome. One caveat is that such variable importance estimates give insight only into the magnitude, not the direction, of the relationship between the predictor variables and the dependent variable (Hall, Phan, & Ambati). We took the top 20 most important variables in each of the six models, which results in a list of 59 unique predictors. Among the 59 unique predictors, there are 16 predictors that appear in the top 20 lists of at least three models, as shown in Table 3. Three themes emerge from the list of the 16 most important variables in the prediction of whether students are low performing in mathematics.

The first theme is home environment support, which includes the home educational resources scale, and variables regarding students' use and possession of computers or tablets. In particular, the home educational resources scale appears in the top 20 lists of all six models.

The second theme is students' attitudes and engagement. Students' confidence in mathematics appears in the top 20 lists across all models, and students' confidence in science appears in three models. Students' educational expectations, their views on engaging teaching, and other student characteristics (including race and age) also play important roles in at least three models in predicting mathematics performance.

School environment is the third theme. Variables including the percentage of students eligible for free or reduced-price lunch, school principals' reporting on school discipline and schools' emphasis on academic success, and total instructional time all contribute significantly to the prediction in at least half of the models.

Conclusions

This study shows how machine learning could be applied to large-scale international studies. It also provides a comparison of six commonly used machine learning models in identifying students with low mathematics performance, which would allow educators to effectively allocate resources to intervene and thus improve their performance. The result shows that logistic regression, elastic net, and extreme gradient boosting all perform well in terms of balanced accuracy and sensitivity. Given that logistic regression is easier to interpret than the other models, it is preferred in this context.

Through analyzing 142 student and school variables, this research confirms the findings in the literature about characteristics that may be associated with mathematics performance, including students' home educational resources and attitudes toward mathematics. It also sheds light on a few variables that receive less research attention, including variables about students' use and possession of computers or tablets.

References

Allison, P. D. (2001). *Missing Data.* Thousand Oask, California: Sage Publications.

Brown-Jeffy, S. (2009). School Effects: Examining the Race Gap in Mathematics Achievement. *Journal of African American Studies*, 388.

Burtless, G. (2011). *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success.* Washington DC: Brookings Institution Press.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., . . . Li, Y. (2018). xgboost: Extreme Gradient Boosting. *R package version 0.71.1.*

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement. *Education Policy Analysis Archives*.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via. *Journal of Statistical Software*.

Gabriel, F., & Jason Signolet, M. W. (2018). A Machine Learning Approach to Investigating The Effects of Mathematics Dispositions on Mathematical Literacy. *International Journal of Research & Method in Education*, 306-327.

García, V., Mollineda, R., & Sánchez, J. (2009). Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions. *Pattern Recognition and Image Analysis*.

Geary, D. C. (2011). Consequences, Characteristics, and Causes of Mathematical Learning Disabilities and Persistent Low Achievement in Mathematics. *Journal of Developmental and Behavioral Pediatrics*, 250–263.

Hall, P., Phan, W., & Ambati, S. S. (n.d.). *Ideas on Interpreting Machine Learning.* Retrieved from O'Reilly Ideas: https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning

Hastie, T., & Qian, J. (2014). *Glmnet Vignette.* Retrieved from https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning.* New York, NY: Springer Publishing Company.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning with Applications in R.* New York, NY: Springer Publishing Company.

Jordan, N. C. (2010). Early Predictors of Mathematics Achievement and Mathematics Learning Difficulties. *Encyclopedia on Early Childhood Development*.

Kuhn, M. (2017). caret: Classification and Regression Training. *R package version 6.0-78*.

Lantz, B. (2013). *Machine Learning with R.* Birmingham: Packt Publishing.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 18-22.

McGraw, R., Lubienski, S. T., & Strutchens, M. E. (2006). A Closer Look at Gender in NAEP Mathematics Achievement and Affect Data: Intersections with Achievement, Race/Ethnicity, and Socioeconomic Status. *Journal for Research in Mathematics Education*, 129-150.

Mullis, I. V., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics.* Boston College.

National Mathematics Advisory Panel. (2008). *Foundations for Success: The Final Report of the National Mathematics Advisory Panel.* Washington, DC: U.S. Department of Education.

Polyzou, A., & Karypis, G. (2016). Grade Prediction with Course and Student Specific Models. In *Advances in Knowledge Discovery and Data Mining* (pp. 89-101). Springer.

Provasnik, S., Malley, L., Stephens, M., Landeros, K., Perkins, R., & and Tang, J. (2016). *Highlights From TIMSS and TIMSS Advanced 2015: Mathematics and Science Achievement of U.S. Students in Grades 4 and 8 and in Advanced Courses at the End of High School in an International Context.* Washington, DC: U.S. Department of Education.

R Core Team. (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 210-229.

Sansone, D. (2017). Beyond Early Warning Indicators: High School Dropout and Machine Learning. *Social Science Research Network*.

Therneau, T., & Atkinson, B. (2018). rpart: Recursive Partitioning and Regression Trees. *R package version 4.1-13*.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S. Fourth Edition.* New York, NY: Springer.

Wang, D. B. (2004). Family background factors and mathematics success: A comparison of Chinese and US students. *International Journal of Educational Research*, 40-54.

Yoo, J. E. (2018). TIMSS 2011 Student and Teacher Predictors for Mathematics Achievement Explored and Identified via Elastic Net. *Frontiers in Psychology*, 317.

Zahedifard, M., Attarzadeh, I., Pazhokhzadeh, H., & Malekzadeh, J. (2015). Prediction of students' performance in high school by data. *International Academic Journal of Science and Engineering*.

Tables and Figures

Figure 1: Study design

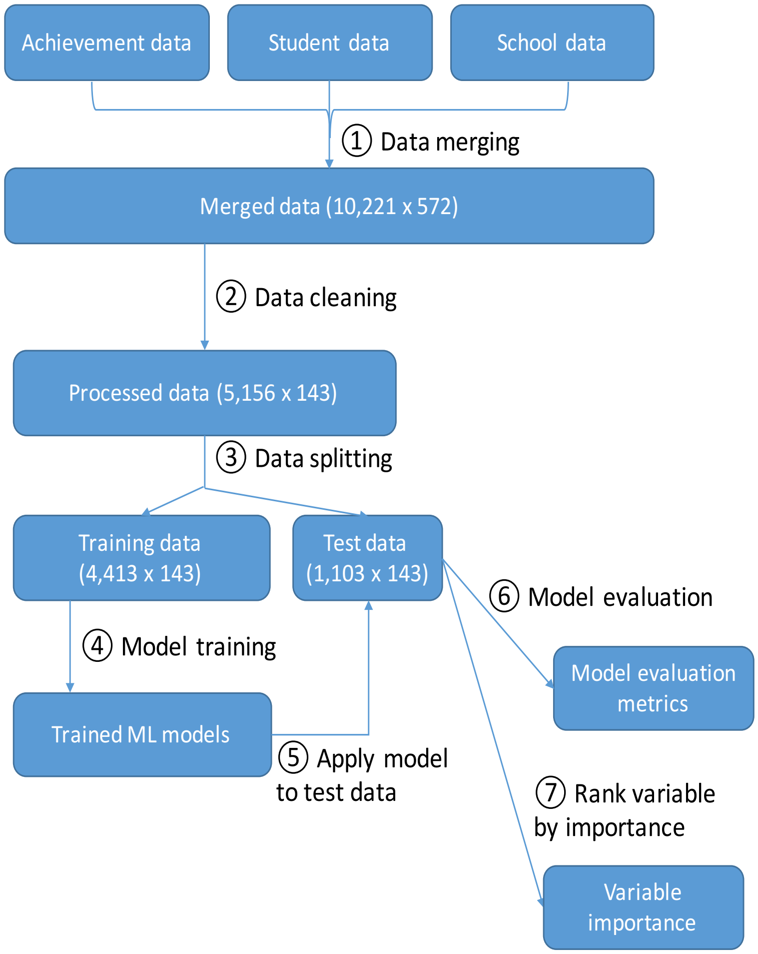

Table 1: Training and test data

	N	**Low-mathematics achievers**	**Non-low-mathematics achievers**
Data	5,516	1,256	4,260
Training data	4,413	1,005	3,408
Test data	1,103	251	852

Table 2: Evaluation measures of machine learning models

Evaluation measure	Equation	Model					
		Logistic regression	Elastic net	Decision tree	Random forest	Extreme gradient boosting	Neural network
Accuracy	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$	85.7%	85.9%	81.5%	83.3%	85.0%	81.0%
Balanced accuracy	$\frac{1}{2}\times\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$	76.0%	75.0%	65.5%	70.0%	74.9%	69.1%
Sensitivity/Recall	$\frac{TP}{TP+FN}$	58.2%	55.0%	36.3%	45.4%	56.6%	47.4%
Specificity	$\frac{TN}{TN+FP}$	93.8%	95.0%	94.8%	94.5%	93.3%	90.8%
Positive predicted value/Precision	$\frac{TP}{TP+FP}$	73.4%	76.2%	67.4%	70.8%	71.4%	60.4%
Negative predicted value	$\frac{TN}{TN+FN}$	88.4%	87.7%	83.5%	85.5%	87.9%	85.4%

Note: TP = number of true positive cases; FP = number of false positive cases; TN = number of true negative cases; FN = number of false negative cases

Table 3: Variables selected by at least three machine learning models

Variable description	Variable source	Appear in the top 20 most important variables						
		Count	Logistic regression	Elastic net	Decision tree	Random forest	Extreme gradient boosting	Neural network
Home Educational Resources Scale	Student questionnaire	6	X	X	X	X	X	X
Students Confident in Mathematics Scale	Student questionnaire	6	X	X	X	X	X	X
[US National Variable] Race/Ethnicity - White, Not Hispanic	Student questionnaire	5	X	X		X	X	X
[US National Variable] Percentage of student elibigble for free or reduced-	School questionnaire	4			X	X	X	X
[US National Variable] How far in your education do you expect to go?	Student questionnaire	4	X		X	X	X	
How often do you use a computer or tablet for schoolwork besides at home or	Student questionnaire	4	X	X	X		X	
Students' Views on Engaging Teaching in Mathematics Lessons Scale	Student questionnaire	4	X		X	X	X	
School Discipline Problems - Principals' Reports Scale	School questionnaire	3			X	X	X	
School Emphasis on Academic Success – Principals' Reports Scale	School questionnaire	3			X	X	X	
Total Instructional Time in Minutes	School questionnaire	3			X	X	X	
Student Age	Student questionnaire	3			X	X	X	
Students Confident in Sciences Scale	Student questionnaire	3	X			X	X	
[US National Variable] Race/Ethnicity -	Student questionnaire	3	X	X				X
[US National Variable] Race/Ethnicity -	Student questionnaire	3		X			X	X
[US National Variable] Have you ever repeated a grade in elementary school?	Student questionnaire	3	X	X				X
[US National Variable] Do you have a computer or tablet that is shared with other people at home?	Student questionnaire	3	X	X				X

Appendix A – Technical Details in Data Preprocessing

The data cleaning process involved the following steps:

1. Create the binary outcome variable, [low_mathematics], which equals 1 if the student is considered low performing in mathematics (and 0 otherwise), defined as having the average of all 5 plausible values for mathematics below the intermediate international benchmark (475).
2. Remove ID-related variables (e.g., [idstud] – student ID), file maintenances variables (e.g., [itdate] – date of testing), weights (e.g., [totwgt] – total student weight), plausible values (e.g., [bsmmat01] – the first plausible value of mathematics), and string variables (e.g. [bsxg03bt] – the U.S.-specific variable about language spoken at home, which is in strings)
3. For a few U.S.-specific variables, remove the international version of such variables because the U.S. versions provide more granular information. For example, for the background question, "how far in education do you expect to go," the U.S. version ([bsng08]) distinguishes between "finish master's degree" and "finish doctorate degree," while the international version ([bsbg08]) only has "finish postgraduate degree."
4. For all derived scale variables (e.g., [bsbgscm] – "Student Confident in Mathematics/SCL"), remove their index versions (e.g., [bsdgscm] – "Student Confident in Mathematics/IDX"), and remove the questionnaire-level variables that these scale variables are derived from (e.g., from [bsbm19a] - "I usually do well in mathematics" to [bsbm19i] – "mathematics makes me confused")
5. Remove variables that have more than 15% missing.[12] The threshold of 15% is chosen to maintain at least half of the original samples after list-wise deletion, which is considered the safest method of handling missing data (Allison, 2001). Increasing the threshold will lead to a smaller sample size after list-wise deletion, while decreasing it will lead to a smaller number of predictors.[13]
6. Dummy-code all binary variables. Convert Liker scale variables (e.g., [bsbg08]) to continuous. For non-ordinal variables (e.g., [msrace2] – race/ethnicity), dummy code each variable level and remove the last level (e.g., the newly coded [msrace2_other]).
7. Construct a correlation matrix among all remaining variables to identify variable pairs that are highly correlated (> 0.95) with each other and remove the variable with the largest mean absolution correlation (e.g., [bcbg07b] "total instructional time/minutes" is highly correlated with [bcdg07hy] – "total instructional hours per year," and the latter is removed).
8. Standardize all variables so that they are on the same scale with a mean of 0 and standard deviation of 1 (Lantz, 2013).

This data cleaning process resulted in 5,516 grade 8 students (54% of the original sample) with 143 student and school variables.

[12] Variable levels such as "Not administered" and "Omitted or invalid" are counted as missing.
[13] In robustness tests where 10% and 20% were used as thresholds, respectively, the findings are consistent with what is documented in the Results and Discussion section.

Appendix B – Variables used in machine learning models

	Variable Name	Variable Label
1	low_mathematics	outcome variable: low-mathematics-achiever
2	bnrgcal1	calculator survey\use of calculator
3	bsbg06a	gen\home possess\computer tablet own
4	bsbg06b	gen\home possess\computer tablet shared
5	bsbg06c	gen\home possess\study desk
6	bsbg06d	gen\home possess\own room
7	bsbg06e	gen\home possess\internet connection
8	bsbg06f	gen\home possess\own mobile phone
9	bsbg06g	gen\home possess\gaming system
10	bsbg06h	gen\home possess\<country specific>
11	bsbg10a	gen\born in <country>
12	bsbg14a	gen\internet use\access textbooks
13	bsbg14b	gen\internet use\access assignments
14	bsbg14c	gen\internet use\collaborate with classmates
15	bsbg14d	gen\internet use\communicate with teacher
16	bsbg14e	gen\internet use\find info to aid in math
17	bsbg14f	gen\internet use\find info to aid in science
18	bcbg08a	gen\have place for schoolwork
19	bcbg09a	gen\student achievement used to assign\math
20	bcbg09b	gen\student achievement used to assign\science
21	bcbg11a	gen\existing science laboratory
22	bcbg11b	gen\existing assistance during exp
23	bcbg12	gen\have school library
24	bcbg17a	gen\use incentives\math
25	bcbg17b	gen\use incentives\science
26	bcbg17c	gen\use incentives\other
27	bcbg22a	gen\degrees in education leadership\isced 7
28	bcbg22b	gen\degrees in education leadership\isced 8
29	bnrgcal2	calculator survey\frequency of using calculator
30	bsbg03	gen\often speak <lang of test> at home
31	bsbg05	gen\digital information devices
32	bsbg11	gen\about how often absent from school
33	bsbg12	gen\how often breakfast on school days
34	bsbg13a	gen\how often use computer tablet\home
35	bsbg13b	gen\how often use computer tablet\school
36	bsbg13c	gen\how often use computer tablet\other
37	bsbs25ab	sci\how often teacher give you homework/science
38	bsbm25ba	math\how many minutes spent on homework/mathematics
39	bsbm26aa	math\extra lessons last 12 month\mathematics

40	bsbs26ab	sci\extra lessons last 12 month\science
41	bsbm26ba	math\extra lessons how many month\mathematics
42	bsbs26bb	sci\extra lessons how many month\science
43	bsbs37a	sci\agree\learn science will help me
44	bsbs37b	sci\agree\need science to learn other things
45	bsbs37c	sci\agree\need science to get into <uni>
46	bsbs37d	sci\agree\need science to get job i want
47	bsbs37e	sci\agree\like job involving science
48	bsbs37f	sci\agree\get ahead in the world
49	bsbs37g	sci\agree\more job opportunities
50	bsbs37h	sci\agree\parents think science important
51	bsbs37i	sci\agree\important to do well in science class
52	bsbm38aa	math\how often tch give you hmwk\mat
53	bsdmwkhw	weekly time spent on math homework
54	bsdswkhs	weekly time spent on science homework
55	bcbg03a	gen\students background\economic disadva
56	bcbg03b	gen\students background\economic affluen
57	bcbg04	gen\percent of students <lang of test>
58	bcbg05a	gen\how many people live in area
59	bcbg05b	gen\immediate area of sch location
60	bcbg06a	gen\free meals\breakfast
61	bcbg06b	gen\free meals\lunch
62	bcbg07c	gen\instructional days in 1 calender week
63	bcbg16a	gen\fill teaching vacancies\math
64	bcbg16b	gen\fill teaching vacancies\science
65	bcbg16c	gen\fill teaching vacancies\other
66	bcbg18a	gen\degree probs teach\arriving late at school
67	bcbg18b	gen\degree probs teach\absenteeism
68	bcbg21	gen\highest level of formal education
69	bcdg03	school composition by std background
70	bcbg07a	gen\instructional days per year
71	bcbg07b	gen\total instructional time\minutes
72	bcbg10	gen\total number computers
73	bcbg19	gen\years principal altogether
74	bcbg20	gen\years principal at this school
75	itsex	sex of students
76	bsdage	students age
77	bsbgher	home educational resources/scl
78	bsbgssb	students sense of school belonging/scl
79	bsbgsb	student bullying/scl
80	bsbgslm	students like learning mathematics/scl

81	bsbgeml	engaging teaching in math lessons/scl
82	bsbgscm	student confident in mathematics/scl
83	bsbgsvm	students value mathematics/scl
84	bsbgsls	students like learning science/scl
85	bsbgesl	engaging teaching in science lessons/scl
86	bsbgscs	student confident in sciences/scl
87	bsbgsvs	students value science/scl
88	bcbgsrs	instr aff by sci res shortage-prncpl/scl
89	bcbgeas	school emph on acad success-prncpl/scl
90	bcbgdas	school discipline problems-prncpl/scl
91	bsng07a.Less.than.high.school	nat\derived\highest lvl edu\mother
92	bsng07a.Some.high.school	nat\derived\highest lvl edu\mother
93	bsng07a.High.school.graduate	nat\derived\highest lvl edu\mother
94	bsng07a.Associate.s.degree..2.year.college.program.	nat\derived\highest lvl edu\mother
95	bsng07a.Bachelor.s.degree..4.year.college.program.	nat\derived\highest lvl edu\mother
96	bsng07a.Master.s.degree.or.professional.degree..MD..DDS..lawyer..minister.	nat\derived\highest lvl edu\mother
97	bsng07a.Doctorate..Ph.D...or.Ed.D..	nat\derived\highest lvl edu\mother
98	bsng07b.Less.than.high.school	nat\derived\highest lvl edu\father
99	bsng07b.Some.high.school	nat\derived\highest lvl edu\father
100	bsng07b.High.school.graduate	nat\derived\highest lvl edu\father
101	bsng07b.Associate.s.degree..2.year.college.program.	nat\derived\highest lvl edu\father
102	bsng07b.Bachelor.s.degree..4.year.college.program.	nat\derived\highest lvl edu\father
103	bsng07b.Master.s.degree.or.professional.degree..MD..DDS..lawyer..minister.	nat\derived\highest lvl edu\father
104	bsng07b.Doctorate..Ph.D...or.Ed.D..	nat\derived\highest lvl edu\father
105	bsbg09a.YES	gen\mother born in the us
106	bsbg09a.NO	gen\mother born in the us
107	bsbg09b.YES	gen\father born in the us
108	bsbg09b.NO	gen\father born in the us

109	msrace2.White..Not.Hispanic	*nat\derived race-collapsed*
110	msrace2.Black..Not.Hispanic	*nat\derived race-collapsed*
111	msrace2.Hispanic	*nat\derived race-collapsed*
112	msrace2.Asian	*nat\derived race-collapsed*
113	msrace2.Two.or.more.races	*nat\derived race-collapsed*
114	bcxg07.Regular.public.school	nat\type of school
115	bcxg07.A.regular.public.school.with.a.magnet.program	nat\type of school
116	bcxg07.A.magnet.school.or.school.with.a.special.program.emphasis	nat\type of school
117	bcxg07.Charter.school	nat\type of school
118	bcxg07.Private..independent.	nat\type of school
119	bcxg07.Private..religiously.affiliated.	nat\type of school
120	bsxg04a	nat\activ outside school\sports team
121	bsxg04b	nat\activ outside school\music instrumnt
122	bsxg04c	nat\activ outside school\other class
123	bsxg04d	nat\activ outside school\club
124	bsxg05a	nat\participated in activ\science fair
125	bsxg05b	nat\participated in activ\science club
126	bsxg05c	nat\participated in activ\science comp
127	bsxg14a	nat\repeat grade\elementary
128	bsxg14b	nat\repeat grade\middle
129	bcxg20b	nat\teacher eval\math\oberv by external persons
130	bcxg20c	nat\teacher eval\math\student achieve
131	bcxg20d	nat\teacher eval\math\teacher peer review
132	bcxg21b	nat\teacher eval\science\oberv by external persons
133	bcxg21c	nat\teacher eval\science\student achieve
134	bcxg21d	nat\teacher eval\science\teacher peer review
135	bsng08	nat\derived\how far educ
136	bsxg13b	nat\days absent last month
137	bsxg30	nat\diff test compared oth tests
138	bsxg31	nat\how hard tried compared oth tests
139	bsxg32	nat\how important to do well
140	bcxg06	nat\percent of students lep/ell
141	bcxg09	nat\avg income lvl of sch immediate area
142	pctfrpl	*nat\percent stud free reduce lunch-categorized*
143	pubpriv	*nat\public private school indicator*