A Comparison of Machine Learning Algorithms for Predicting Student Performance in an Online Mathematics Game

Paper submitted to the American Educational Research Association Annual
Meeting San Diego, California
April 2022

Ji-Eun Lee Amisha Jindal Sanika Nitin Patki

Ashish Gurung Reilly Norum Erin Ottmar

Lee, J. E, Jindal, A., Patki, S., Norum.,R., & Ottmar, E. (2022). A comparison of machine learning algorithms for predicting student performance in an online mathematics game. Full paper presented in the 2022 Annual meeting of the American Educational Research Association (AERA). San Diego, CA.

Abstract

This paper demonstrates how to apply Machine Learning (ML) techniques to analyze student interaction data collected in an online mathematics game. We examined 1) how different ML algorithms influenced the precision of middle-school students' (*N* = 359) performance prediction and 2) what types of in-game features were associated with student math knowledge scores. The results indicated that the Random Forest algorithm showed the best performance in predicting posttest math knowledge scores among the seven algorithms employed. Out of 37 features included in the model, the validity of the students' first mathematical transformation was the most predictive of their math knowledge scores. Implications for game learning analytics and supporting students' algebraic learning are discussed based on the findings.

Keywords

Mathematics Learning, Online Mathematics Game, Prediction, Random Forest

Introduction

In recent years, there has been significant growth of digital game-based learning. In turn, *game learning analytics* has received a lot of research attention. Due to their highly interactive environments compared to other types of educational technologies, digital learning games record tremendous quantities of student actions in the form of log files (Serrano-Laguna et al., 2014). By leveraging these data, game learning analytics have provided information and insights into student in-game behaviors, the effectiveness of educational games, and the improvement of game design (Alonso-Fernandez et al., 2019).

However, many studies in the field have focused on using simple aggregations of student actions or the correctness of students' answers, and relatively little attention has been paid to qualitative aspects of students' behaviors. Moreover, while it is encouraged to compare various machine learning techniques to draw a better prediction/classification result, much of the research uses only one or two simple methods rather than choosing the best model after evaluating multiple techniques.

Over the past several years, our team has developed an online mathematics game that aims to improve students' algebraic understanding. Our previous studies have revealed that the game is effective in improving students' mathematical understanding (Authors, 2019; 2021). However, the best feature sets or algorithms to predict student math performance have not yet been identified, warranting further investigation. Hence, extending our prior work, we examine how different features provided for optimizing machine learning techniques influence the precision of predicting middle-school students' mathematics performance. In particular, we use simple aggregations of students' actions and hand-coded data that measures the qualitative aspects of students' actions in the game. We address the following research questions:

1. Which machine learning algorithm provides the best results for students' posttest math knowledge scores prediction?
2. What kinds of student behaviors in the game are associated with students' math performance?

Literature Review

Machine Learning Algorithms for Prediction

The most commonly used supervised learning algorithms in game learning analytics are linear/logistic regression, decision trees, and Support Vector Machines (SVM) (Alonso Fernandez et al., 2019). Each algorithm has its pros and cons; for example, logistic regression is relatively easy to implement and does not require feature scaling; however, it tends to show poor performance on non-linear data. In contrast, Random Forest (RF) produces good performance on imbalanced datasets and shows good handling of large datasets or missing values. SVM is well suited for non-linear data such as image data or data that has many features. One of the challenges in game learning analytics is that variables extracted from the data are often game specific, which makes it challenging to build a model that can be generalized across contexts (Serrano-Laguna et al., 2014). Thus, it is important to compare various techniques to identify the algorithm that suits the data best as well as to improve the accuracy of models.

Factors Affecting Students' Performance in Online Learning Games

Studies have found that students' in-game behaviors and prior knowledge are predictive of their learning outcomes. For instance, one study (Nguyen et al., 2020) investigated factors related to middle school students' decimal understanding in a digital mathematics learning game. Among the features included in the model, students' pretest scores and two in-game behaviors showed significant and positive associations with their posttest scores. Data other than the game logs were not predictive of posttest scores but were predictive of game enjoyment. Another study (Shute et al., 2015) examined relationships among middle school students' prior knowledge, in-game progress, persistence, and their understanding of physics using the digital learning game. Similar to Nguyen et al.'s study, students' pretest scores and in-game progress both significantly predicted their understanding of physics.

As such, although simple aggregations of students' actions or interactions in the game are useful to predict learning outcomes, adding more data or additional information may improve prediction accuracy. Thus, this study builds prediction models using both simple aggregations of students' actions and qualitative, hand-coded data of the students' game exploration strategies.

Method

Game Description

We used data collected in a web-based digital learning game that helps middle-school students' conceptual and procedural understanding of algebra. In this game, numbers and mathematical symbols are reified as movable objects so that students can dynamically manipulate and transform numbers or mathematical expressions on the screen. The game consists of 252 problems that cover a variety of mathematical concepts presented in order of increasing complexity. Each problem consists of two mathematically equivalent mathematical expressions: a start state (121×144) and a goal state (11×132×12) (Figure 1). The goal of each problem is to transform the start state into a target goal using permissible gesture actions. Students can also reset the expression to the initial state, request hints, and reattempt the problems as many times as they want. Our prior work has shown that the game is effective in improving students' algebraic understanding after controlling for their prior knowledge (Author, 2019; 2021).

Participants and Research Procedure

We used data collected from a larger randomized control trial conducted in Fall 2019, which consisted of 359 sixth and seventh-grade students from six middle schools located in the Southeastern U.S. Of the 359 participants (male: 51%, female: 39%, not identified: 10%), the majority of the students were in 6th grade (85%). The students took a pretest of their understanding of algebra and math anxiety before the intervention. After that, they played the game individually at their own pace for four 30-minute intervention sessions during the regular math classes. On average, the students solved 97.4 distinct problems in the game (*SD* = 34.2). After the intervention, they took a posttest measuring math knowledge and math anxiety with items similar to those given at the pretest.

Data Pre-processing

We extracted log data from the database and created simple aggregations of log files using JavaScript, which resulted in 37 features on each problem. Among 37 preliminary features, we eliminated irrelevant features and constructed 32 variables of student behavior. Because the students played the game individually at their own pace, not all problems were attempted by the students, which led to a lot of missing values. In order to minimize the amount of missing data, we selected the problems that were attempted by at least 150 students, which resulted in a subset of 98 problems. In addition, we detected the outliers and masked them for further statistical computations. Figure 2 summarizes our data preprocessing and evaluation process.

Measures

In addition to the in-game variables, we added three variables collected through the assessment to predict posttest math knowledge scores: pretest math knowledge scores, pretest math anxiety scores, and posttest math anxiety scores. Both pretest and posttest math knowledge scores were measured using 11 items selected from two validated measures (Rittle-Johnson et al., 2011; Star et al., 2015). The KR-20 coefficients of these 11 items were 0.69 at pretest and 0.76 at posttest. To measure students' math anxiety, we used 13 items adapted from an established measure (Ganley & McGraw, 2016). The Cronbach's �� of these items were 0.87 at pretest and 0.91 at posttest.

Furthermore, to assess the qualitative aspects of students' problem-solving in the game, we included three additional variables: mathematical expressions made by students, mathematical strategies, and the productivity of these strategies (hereafter, productivity), which refers to whether or not the student's action moved them closer to the goal state of the problem. We hand-coded the mathematical strategies and the productivity using the log data collected in the game. The intra correlation coefficients of the hand-coding ranged between 0.91-0.98 for math strategies and 0.74-0.96 for productivity. Table 1 lists the three types of variables, examples, and descriptions of each feature included in the prediction models.

Data Analyses

In order to examine which algorithms best predict students' performance (RQ1), we compared seven different machine learning algorithms: RF Regressor, Multilayer Perceptron (MLP), AdaBoost, Linear Lasso, Logistic Regression, Bagging Regressor, and Support Vector Regressor (SVR). We used Mean Square Error (MSE), Mean Absolute Error (MAE), and R^2 to evaluate the performance of the seven algorithms. All data analyses were performed using packages in Python version 3.7.

Results

Correlation Analysis

We conducted a Pearson correlation analysis for the exploration of the data on students' behavioral features, assessment features, and math knowledge scores (Figure 3). Results indicated that pretest math knowledge scores had an extremely high correlation with the posttest scores; thus, we eliminated pretest math knowledge scores from further analysis. Four student behavior variables -- the trial of the problem (i.e., whether or not the student tried the problem; labeled "tried" in Figure 3), the number of clovers (i.e., rewards in the game) earned in the first attempt ("clover_first"), the validity of the first mathematical transformation on the first attempt (i.e., whether or not the first action on the first attempt is a valid step; "interaction_step_first"), and the validity of the first step on the last attempt ("interaction_step_last") -- showed positive and higher correlations ($r \geq 0.5$) with posttest scores than other features.

RQ1: Comparison of Machine Learning Algorithms for Student Performance Prediction

We examined how different features provided for the optimization of seven machine learning techniques influence the precision of students' posttest scores prediction. As shown in Table 2, RF showed the lowest values for MSE (2.858) and MAE (1.354), while MLP had the highest values of MSE (20.289) and MAE (2.405) among seven ML algorithms. Regarding prediction accuracy, the model with the RF algorithm showed the highest accuracy score among seven models, explaining 40.8% of the variance in posttest scores. Similar to the results of error metrics, MLP indicated the lowest accuracy score. Together, the results indicated that RF showed the best performance in predicting student posttest math knowledge scores, while MLP showed the lowest accuracy and highest error values for our dataset.

RQ2: Students' Math Performance Prediction

We examined the relations between student features and their posttest scores using the prediction model with RF, as it outperformed the other six ML algorithms. As shown in Figure 4, the most influential feature in predicting the posttest scores was the "interaction_step_first," which indicates whether or not a student made a mathematically valid transformation on their first problem-solving. In other words, if a student made a valid first step without making any errors or random clicks, the student was more likely to receive a higher posttest score. The second most influential feature was "use_hint" which refers to whether or not the students requested a hint. Although the use of hints was one of the important features in predicting the posttest scores, the direction of the association was negative. Specifically, the students who requested hints more frequently tended to achieve lower posttest scores. Lastly, while the two student assessment features showed relatively higher importance values, the three features of math strategies had relatively lower importance in predicting the posttest scores.

Conclusion

In this paper, we examined how different features provided for the optimization of machine learning techniques influence the precision of predicting middle-school students' math knowledge scores, using student interaction data collected from an online mathematics game. First, we examined how different ML algorithms influenced the precision of students' math knowledge scores prediction. The results revealed that the RF outperformed the other six algorithms for our dataset. The results confirm that the RF algorithm performs well on the imbalance dataset as many of the features included in the prediction models were positively skewed. A possible explanation for the worst performance of MLP may be due to the relatively small amount of data. In addition, our results indicated the effectiveness of the prediction model in estimating student posttest scores, even after excluding the pretest scores that had a very high correlation with the posttest scores.

Second, we investigated what types of features in the game were associated with student posttest math knowledge scores. The prediction model with the RF algorithm indicated that the validity of students' first mathematical transformation on their first attempt was the most influential predictor for math performance, which implies that noticing the pattern or structure of the problems may play an important role in students' mathematical understanding, rather than rushing into problem-solving.

Acknowledgements

We would like to thank the teachers and students for their participation; members of the Math, Abstraction, Play, Learning, and Embodiment (MAPLE) Lab for their work; Erik Weitnauer and members of Graspable Math Inc. for programming support; and Neil Heffernan and the ASSISTments Team for their support. This work was supported by Institute of Education Sciences: [Grant Number R305A180401] and National Science Foundation: [Grant Number 2142984]. Erin Ottmar was a designer and co-developer of From Here to There! The other authors declare no conflicts of interest associated with the publication of this manuscript.

References

Authors (2019)

Authors (2021)

Alonso-Fernandez, C., Calvo-Morata, A., Freire, M., Martinez-Ortiz, I., & Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: A systematic literature review. *Computers & Education*, *141*, 103612.

Ganley, C. M., & McGraw, A. L. (2016). The development and validation of a revised version of the math anxiety scale for young children. *Frontiers in Psychology*, *7*, 1181.

Nguyen, H. A., Hou, X., Stamper, J., & McLaren, B. M. (2020). Moving beyond test scores: Analyzing the effectiveness of a digital learning game through learning analytics. In A. N. Rafferty, J. Whitehill, C. Romero, & V. Cavalli-Sforza (Eds.)*, Proceedings of the 13th International Conference on Educational Data Mining* (pp. 487-495). International Educational Data Mining Society.

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct-modeling approach. *Journal of Educational Psychology*, *103*(1), 85.

Serrano-Laguna, Á., Torrente, J., Moreno-Ger, P., & Fernández-Manjón, B. (2014). Application of learning analytics in educational videogames. *Entertainment Computing*, *5*(4), 313-322.

Shute, V. J., D'Mello, S., Baker, R., Cho, K., Bosch, N., Ocumpaugh, J., ... & Almeda, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education*, *86*, 224-235.

Star, J. R., Pollack, C., Durkin, K., Rittle-Johnson, B., Lynch, K., Newton, K., & Gogolen, C. (2015). Learning from comparison in algebra. *Contemporary Educational Psychology*, *40*, 41-54.

Table 1

Features Included in the Final Prediction Models

Type of features	Examples of features Number of features	
Student behavior (extracted from the log data) number of steps made;	number of visits; number of errors made; number of clovers received; number of reattempts;	number of resets; hint usage; time taken to solve each problem 32

Student assessment	pretest math knowledge scores; posttest math anxiety scores; pretest math anxiety scores	3
Math strategies	mathematical expressions made by a student; math strategies (e.g., calculating, decomposing, commuting); productivity of mathematical transformation (i.e., whether or not the student’s action moved them closer to the goal state of the problem)	3

Table 2

Prediction Error Measures and Accuracy of Models Predicting Student Math Knowledge Scores

Algorithms	MSE	MAE	R^2
RF	2.858	1.354	0.408

Bagging Regressor	2.968	1.366	0.385
AdaBoost	3.174	1.438	0.343
SVM	3.466	1.475	0.282
Linear Lasso	3.606	1.517	0.253
Logistic	4.937	1.770	-0.021
MLP	20.289	2.405	-3.197

Note: R-squared is negative only when the model does not follow the trend of the data, so fits worse than a horizontal line.

Figure 1

A Sample Problem and Students' Actions in the Game

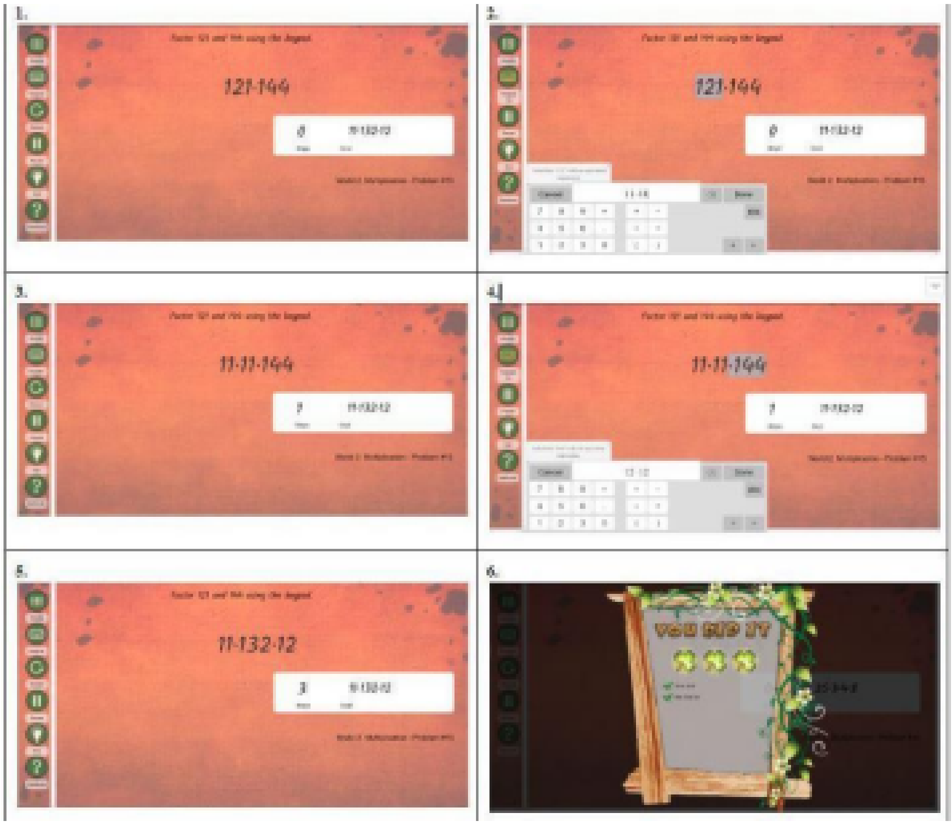


Figure 2

Data Preprocessing and Evaluation Process

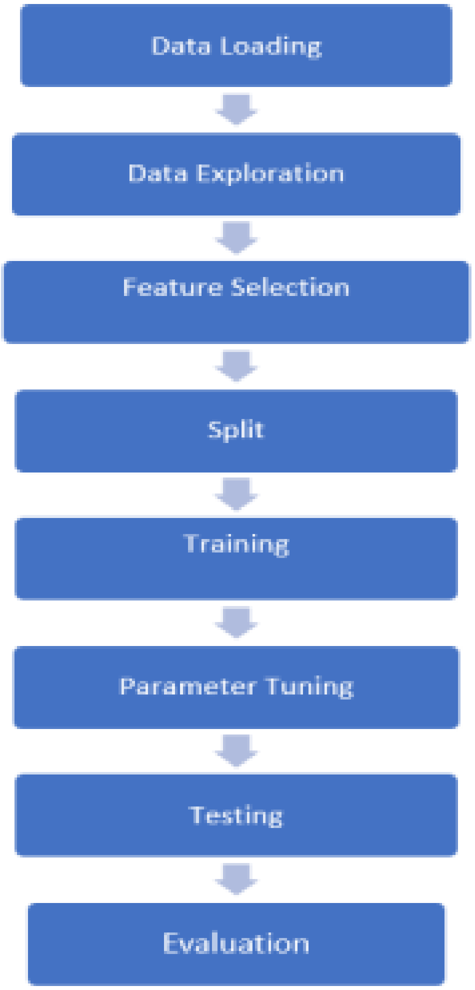
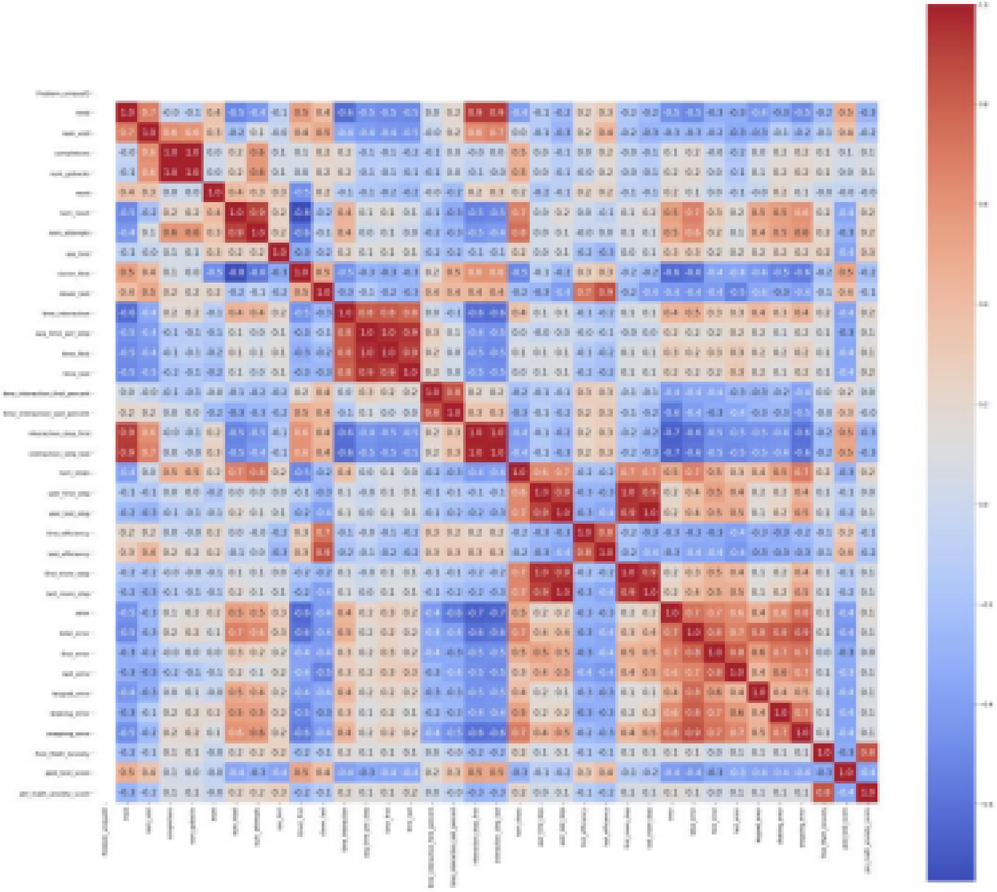


Figure 3

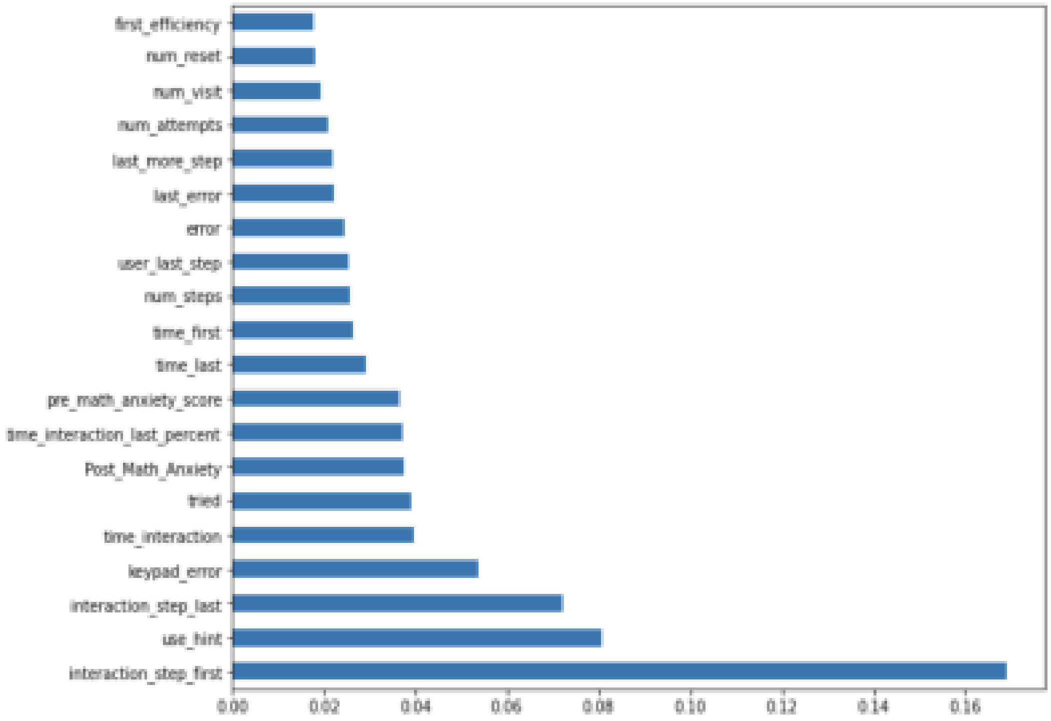
Correlations Among Students' Behavioral Features, Assessment features, and Posttest math Knowledge Scores (for full image: https://tinyurl.com/237z4lfc)



Note: In the correlation matrix, a darker red indicates a stronger positive coefficient, and a darker blue represents a stronger negative coefficient.

Figure 4

The Results of the Random Forest Prediction Model (Feature Importance)



Note: The result represents relative variable importance in the RF regressor for student posttest scores in ascending order.