An Application of Data Mining Methods on In-Game Behaviors: Predicting Student Math Performance

Siddhartha Pradhan, Ji-Eun Lee, Alena Egorova, Janette Jerusal, Erin Ottmar
sppradhan@wpi.edu, jlee13@wpi.edu, aegorova@wpi.edu, jjerusal@wpi.edu, erottmar@wpi.edu
Worcester Polytechnic Institute

Abstract: This study tested the applicability of the Random Forest (RF) prediction model built in Lee et al. (2023) to a different sample with a larger number of students (*N* = 681). Specifically, we predicted middle school students' math knowledge scores using both in-game behavioral features in a digital algebraic learning game and math anxiety scores. Further, we conducted a Principal Component Analysis (PCA) to test whether dimensionality reduction improved the performance of the prediction model. The results showed that the RF regressor performed well in predicting posttest math knowledge scores even when trained on data without pretest scores. Out of the 20 features included in the prediction model, the average number of steps a student takes in their first attempt was the most predictive of their posttest scores. Finally, using PCA as preprocessing to minimize multicollinearity and noise improved the model performance.

Introduction

With the acceleration of the use of digital learning, ensuring the quality of digital learning tools has become more important than ever (Esfijani, 2018). One effective approach to ensure the quality of digital learning tools is to use the learning process and behavioral data collected within these tools to build a *prediction model* for student performance (Qiu et al., 2022). Building a student performance prediction model can provide useful information for teachers and researchers on what student behaviors lead to better performance, which in turn reduces the risk of failing students, and provides a personalized learning environment. While it is encouraged to compare various machine learning techniques to build a model with a better prediction result, much of the research tends to use one or two simple techniques, rather than choosing the best model after evaluating multiple algorithms (Zhang et al., 2021).

Our research team previously applied machine learning techniques to analyze student interaction data in *From Here to There!* (FH2T), a digital algebraic learning game (Lee et al., 2023). We compared the performance of seven machine learning algorithms in predicting students' math knowledge scores and examined what in-game behaviors were associated with students' math knowledge scores. The results indicated that Random Forest (RF) showed the best performance in prediction, and the validity of the students' first mathematical transformation was the most predictive of their math knowledge scores. In this study, we used the methods in our prior work on a different sample with a larger number of students to examine if this prediction model with the RF algorithm is still applicable to other populations. Further, we conduct a Principal Component Analysis (PCA) to test if the dimensionality reduction improves the model performance. We address the following Research Questions:

- RQ1: How effective is the Random Forest (RF) model in predicting posttest math knowledge scores?
- RQ2: Does dimensionality reduction using Principal Component Analysis (PCA) improve the performance of the prediction model?

Background: In-game behaviors and math performance in digital games

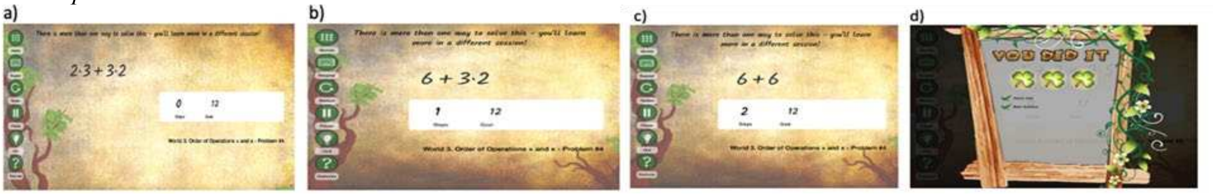
In digital learning game environments, students' in-game behaviors can be used to assess their progress and to provide reliable estimates of learning. For instance, Alonso-Fernández et al. (2020) showed that the prediction model based solely on 19 in-game behaviors had a comparable prediction accuracy to the model with the pretest scores. Regarding the association between in-game behaviors and student performance, studies have found that students' propensity to replay problems (Liu et al., 2017), access to in-game supports (Shute et al., 2021), frequency of exploratory play (Pradhan et al., 2024), significantly predicted post-test scores. As such, these studies found different results as behavioral indicators are often specific to the games and not standardized. Further, few studies in the field consider the inherent correlation among online learning behaviors (Qiu et al., 2022). Thus, this study extends our prior work to build a more generalizable model to predict student performance and conduct PCA to minimize multicollinearity and reduce dimensionality.

Methodology

Game description

FH2T (graspablemath.com/projects/fh2t) is a digital mathematics learning game to improve students' algebraic knowledge and mathematical flexibility (Ottmar et al., 2015). In this game, students must transform a start state (e.g., 2*3+3*2 in Fig 1a) into a mathematically equivalent but perceptually different goal state (e.g., 12 in Fig 1d) using permissible gesture actions (e.g. Fig 1b and Fig 1c). The numbers and mathematical symbols in the game are made into physical objects that can be virtually grasped and moved to supplement the written problems, which allows students to have more in-depth, dynamic learning experiences. Our prior studies have shown that the game is effective in improving students' mathematical knowledge (e.g., Decker-Woodrow et al., 2023).

Figure 1
A Sample Problem and Student's Actions in From Here to There!



Sample

The participants were 7th-grade students from 11 schools in the Southeastern U.S. Of the 1,649 students, 681 students who completed both pre and post-tests were included in the further analyses (52.13% male). A majority of students were White (51.39%), followed by Asian (23.79%), Hispanic (16.74%), and other race/ethnicity (8.08%). On average, the students completed 111.06 problems (*SD* = 55.05) out of 252 problems.

Data-preprocessing and measures

Log and aggregated data were extracted from the FH2T database (osf.io/r3nf2). We constructed an analytical dataset at the student-problem level, which had a total of 42 features. Of the 42 features, 38 features represented students' in-game behaviors, two features assessed the qualitative aspects of student mathematical problem-solving (i.e., the productivity of the first mathematical transformation [hereafter, productivity] on the first and last completed attempt), and two additional features were collected through the self-report assessment (i.e., pre- and post-math anxiety scores). To account for problem-level variation, problem-level features were standardized using Z-scores. We then removed outliers using an interquartile range approach to improve the accuracy of the models. For RQ1, feature selection was performed to improve the performance of the models and reduce computational costs in modeling. The features with correlation coefficients with post-test scores greater than 0.9 and features with variance lower than 0.05 were removed. This resulted in a final analytical dataset of 20 features (see tiny.cc/islsfeaturelist). The data used to train the models was aggregated at the student level by taking the mean of each feature across problems for each student. For RQ2, we used the full 42 features to preserve as much of the variation as possible.

Data analysis

For RQ1, we used the RF Regressor to predict posttest scores using in-game behaviors and math anxiety features. We used three metrics to evaluate the performance of our models: R-square, MAE, and MSE. The data analysis was performed in Python 3.11. The data was split into training data (75%), and testing data (25%) for evaluation. The RF regressor's hyperparameters were tuned using a 5-fold cross-validation in the training data. For RQ2, PCA was performed on the training data after standardization and outlier detection to eliminate multicollinearity and assess whether the performance of the model improved. The optimal number of Principal Components (PC) was chosen by using a 5-fold cross-validation that resulted in the lowest MSE.

Results

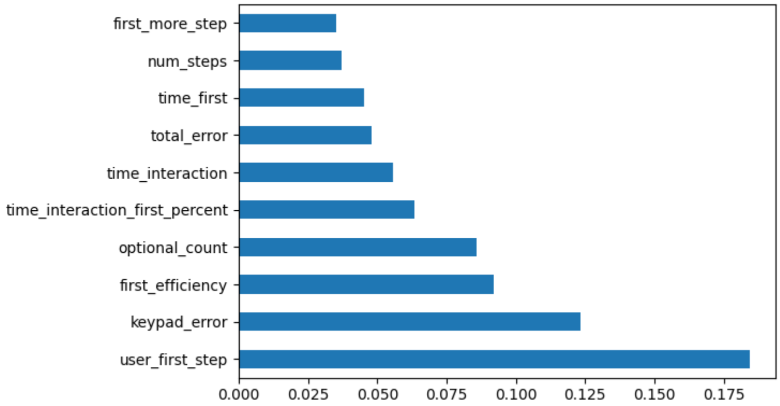
RQ1: Predicting posttest scores using in-game behaviors and assessment features

The RF model had an out-of-sample R^2 of 0.455 (Adjusted R^2 = 0.435), a MSE of 4.61, and a MAE of 1.71 on the testing set. Figure 2 shows the importance of variables in ascending order, where the x-axis represents the Gini importance. The most important feature in predicting posttest scores was the average number of steps that a student took in the first attempt (i.e., 'user_first_step'), which had a negative association with posttest scores.

Subsequently, the number of keypad errors (i.e., calculation errors) was negatively correlated with posttest scores, implying that students who made more keypad errors tended to perform worse in their posttest. Conversely, features that exhibited a positive correlation with posttest scores, labeled 'first efficiency'—signifying the average strategy efficiency of a student's initial transformation in their first attempt—and 'optional count,' representing the total count of optional problems attempted, were also of substantial importance as indicated by the RF model.

Figure 2

Top 10 Important Features Based on the RF Model



RQ2: Predicting post-test performance with Principal Components

The optimal number of principal components (PCs) was determined using cross-validation. Using eight PCs resulted in the lowest cross-validated MSE of 4.74. Thus, we conducted a PCA using eight components (Table 1). The eight PCs altogether explained 82.7% of the variation in the training data. The RF model had an out-of-sample (testing) R^2 of 0.535 (Adjusted R^2 = 0.528). In other words, the predictors explained 53.5% of the variation in the posttest, which showed a higher variance explained than the RF model without dimensionality reduction. Furthermore, this model had an MSE of 4.05 and MAE of 1.58 on the testing set.

Table 1

PCA Loading for the Features in FH2T

Feature	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Variance Explained	0.249	0.166	0.125	0.111	0.058	0.044	0.042	0.032
clover_last	-0.300	-	-	-	-	-	-	-
clover_first	-0.387	-	-	-	-	-	-	-
user_first_step	-	0.400	-	-	-	-	-	-
user_last_step	-	0.398	-	-	-	-	-	-
shaking_error	-	-	0.317	-	-	-	-	-
number_of_steps	-	-	0.316	-	-	-	-	-
post_MA_score*	-	-0.450	-	0.540	-	-	-	-
pre_MA_score*	-	-0.393	-	0.522	-	-	-	-
first_productivity*	-	-	-	-	-0.482	-	-0.356	-
last_productivity	-	-	-	-	-0.493	-	-	-
time_first_interaction	-	-	-	-	-	0.561	-	-
time_interaction_last	-	-	-	-	-	0.459	-	-
number_of_gobacks	-	-	-	-	-	-	0.323	-
optional_count*	-	0.324	-	-	0.447	-	-	0.631
use_hint*	-	-	-	-	-0.352	-	-	0.510

Note: Factor loadings < .3 are suppressed. *Items cross-loaded onto multiple factors.

Discussion

In this paper, we examined whether our methods from the previous study (Lee et al., 2023) were applicable to a different sample with a larger number of students. For RQ1, we examined the performance of the RF regressor identified from the prior study on a different sample at the student level. The prediction model without pretest scores explained a substantial amount of variance in posttest scores, which supports previous studies (Alonso-Fernández et al., 2020). Unlike the previous study, the average number of steps a student takes in their first attempt at a problem was the most influential predictor of posttest scores, and the strategy efficiency score of students' initial transformation in their first attempt was the most influential positive predictor of posttest scores. Together, the results suggest that students who use more efficient problem-solving strategies (i.e., fewer steps/computations to reach the goal state) across problems in the game were more likely to receive a higher posttest math knowledge score. It is plausible that the different results between the prior work and the present study may be due to the different demographic characteristics of the two studies. For RQ2, we used PCA as a preprocessing step for dimensionality reduction to minimize multicollinearity, instead of dropping highly correlated and low-variance features. The results indicated that using PCA to preprocess the data improved overall model performance. This suggests that using the PCs rather than the features led to less overfitting and a more generalizable model. There is a notable decrease in prediction error, and depending on the use case is a viable preprocessing step to improve the robustness and precision of predictive models. Finally, this study did not account for variations that might arise due to demographic features such as gender, race, and socioeconomic status. Future work should focus on assessing the generalizability of the models across demographic features to ensure the results are fair and applicable to all students in the sample.

References

Alonso-Fernández, C., Martínez-Ortiz, I., Caballero, R., Freire, M., & Fernández-Manjón, B. (2020). Predicting students' knowledge after playing a serious game based on learning analytics data: A case study. Journal of Computer Assisted Learning, 36(3), 350-358.

Bonett, D. G. (2012). Replication-extension studies. Current Directions in Psychological Science, 21(6), 409-412.

Decker-Woodrow, L., Mason, C. A., Lee, J. E., Chan. J. Y. C., Sales, A., Liu, A., & Tu, S. (2023). The impacts of three educational technologies on algebraic understanding in the context of COVID-19. AERA Open. 9(1), 1-17.

Esfijani, A. (2018). Measuring quality in online education: A meta-synthesis. American Journal of Distance Education, 32(1), 57-73.

Lee, J. E., Jindal, A., Patki, S. N., Gurung, A., Norum, R., & Ottmar, E. (2023). A comparison of machine learning algorithms for predicting student performance in an online mathematics game. Interactive Learning Environments, 1-15.

Lee, J. E., Chan, J. Y. C., Botelho, A., & Ottmar, E. (2022). Does slow and steady win the race?: Clustering patterns of students' behaviors in an interactive online mathematics game. Educational Technology Research and Development, 70(5), 1575-1599.

Liu, Z., Cody, C., Barnes, T., Lynch, C., & Rutherford, T. (2017). The Antecedents of and Associations with Elective Replay in an Educational Game: Is Replay Worth It? International Educational Data Mining Society.

Ottmar, E., Landy, D., Weitnauer, E., & Goldstone, R. (2015). Graspable mathematics: Using perceptual learning technology to discover algebraic notation. In Integrating touch-enabled and mobile devices into contemporary mathematics education (pp. 24-48). IGI Global.

Pradhan, S., Gurung, A., Ottmar, E. (2024, March). Gamification and Deadending: Unpacking Performance Impacts in Algebraic Learning. In LAK24: 14th International Learning Analytics and Knowledge Conference.

Shute, V., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C. P., ... & Sun, C. (2021). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. Journal of Computer Assisted Learning, 37(1), 127-141.

Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., ... & Chen, P. K. (2022). Predicting students' performance in e-learning using learning process and behaviour data. Scientific Reports, 12(1), 453.

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: method review and comparison analysis. Frontiers in Psychology, 12, 698490.

Acknowledgment

This work was supported by National Science Foundation: [Grant Number 2142984] and Institute of Education Sciences: [Grant Number R305A180401].