International Society of the Learning Sciences

Predicting Learning Gains in an Educational Game using Feature Engineering and Machine Learning

Seyedahmad Rahimi, University of Florida, srahimi@ufl.edu
Curt Fulwider, Florida State University, gcf16@fsu.edu
Shiyan Jiang, North Carolina State University, sjiang24@ncsu.edu
Valerie J. Shute, Florida State University, vshute@fsu.edu

Abstract: In this study, we use feature engineering and machine learning (ML) to develop a regression model to predict students' learning gain while playing an educational game. Specifically, this study used the log data from 199 students' gameplay to build a logistic regression model. The model we created includes 14 features with an accuracy of .61.

Introduction

Educational games are complex environments that can assess students' learning process and outcome. As students interact with the game, they produce rich data which provides evidence of learning. Using machine learning (ML) techniques, we can make use of game data to derive models that predict students' experience. By doing so, we can create educational games that will respond—providing supports or adaptive challenges—to student's learning needs. In this study, we use feature engineering and machine learning to develop a regression model to predict students' learning gain (pre to posttest) using gameplay data from *Physics Playground* (*PP*; Shute et al., 2019).

Method

We examined a dataset (n = 199) from 9th to11th-grade students (*Male* = 104, *Female* = 91, *Other* = 4) in a large public school in the southeastern United States (see Shute et al., 2020). Students self-reported a diverse range of ethnicities: White (42%), Black or African American (31%), Other (19%), Hispanic (8%). Collectively, these students generated about 6 million data points in 5 days of gameplay (about 1.5 hrs of gameplay per day). *PP* is a 2D computer-based game (including 81 levels) created to help secondary school students learn Newtonian physics. Using simple physics agents (e.g., ramp, lever) or manipulating physics parameters (e.g., mass of the ball), students hit a red balloon using a green ball in *PP*'s levels. To access the learning supports, students click on the help button, which is always available during gameplay. The game's incentive system includes earning money resources and spending money opportunities (e.g., buying background music, background image, new ball, or a solution video). We collected log data of students' interactions with the game. The log file contains nearly 17,500 events representing each level played by all 199 participants. 13,312 events were used for modeling after removing missing values. A set of features was distilled based on our prior work on developing evidence-based assessments (Shute et al., 2020). We originally considered 45 features that show students' actions in each game level, such as the coins students earned (gold or silver) and how long students played a level. Our refinement and selection of the feature space was guided by the theory of game-based learning (Plass et al., 2015) and literature about predicting learning gain with interactions with technologies (Crossley et al., 2020). After several rounds of feature selection based on training data for reducing overfitting, we selected 14 features (Table 1) for modeling learning gains using a logistic regression model. The learning gains were measured through two isomorphic tests with multiple-choice items (pretest = 18 items, α = .77; posttest = 18 items, α = .82). The higher the absolute value of a feature weight in a regression model, the stronger its influence on learning gains. All features were normalized using z-score normalization. We tuned the regression model to achieve high accuracy in predicting learning gains.

Results

Our final model achieves a cross-validated accuracy of 0.61 and a cross-validated kappa of 0.12. The features explain 5.31% of the variance of those who had learning gains. The confusion matrix shows that the model had the majority error cases in predicting events that should be false but were predicted as true. This indicates that in the future, we still need to revise the feature space so that the model could identify patterns for positive learning gains. As shown in Table 1, game levels with challenging physics concepts (concept difficulty), learning support, game mechanics difficulty (GM), and agents drawn are good predictors for learning gain while not solving the game level (solved), money earned, number of help button usage, game supports viewed, and number of physics animations viewed negatively predicted learning gains.

Table 1
Definitions of Features and their Weights

Feature	Weight	Definition
Performance		
Time	-0.001	The time, in seconds, a player spends in the level.
Solved[1]	0.720	Whether the player solved the level or not (yes/no).
Money earned	-0.030	The amount of money earned playing the level.
Wallet balance	-0.001	The amount of money in the player's wallet at the end of a level.
Support		
Help button usage	-0.090	The total number of times the help button was used within the level.
Learning support viewed	0.080	The total number of learning supports, of all types, viewed in a level.
Game support viewed	-0.060	The total number of game supports, of all types, viewed in a level.
Physics support viewed	-0.060	The total number of physics supports, viewed in a level.
Physics animations (PA) viewed	-0.010	The total number of physics animations viewed within a level.
Engagement		
Agents drawn	0.020	The number of mechanical agents (e.g., ramp) drawn within a level.
Objects drawn	0.007	The total number of objects (excluding agents) drawn in a level.
Level Traits		
Type[2]	0.003	The type of level. Either manipulation or sketching.
Concept difficulty[3]	0.200	The difficulty of physics concepts included in the level.
Game mechanics (GM) difficulty	0.070	The difficulty of game mechanics included in the level.

[1] Compared to unsolved; [2] Focus is sketching type levels; [3] Focus is hard difficulty

Discussion

Looking at the features contributing to positive learning gain, we can see variables that are related to game performance. Specifically, solving the game levels and drawing agents (simple machines). Other than game performance variables, learning support-related variables are also good predictors of learning gain. That is, the total number of learning supports viewed was a positive predictor of learning gains. Variables that predicted learning loss included variables other than gameplay (e.g., clicking the help button, money earned, game support, and physics animations viewed). These variables indicate that students needed help and showed that students spent time doing tasks which took them out of gameplay. Finally, variables related to a game level's difficulty were also important predictors of learning gains. Specifically, the concept difficulty was a strange predictor compared to game mechanics difficulty. This indicates that difficult concepts and difficult levels in terms of game mechanics contribute more to students' learning. The model we came up with can be used for early detection of learning gain.

References

Crossley, S. A., Karumbaiah, S., Ocumpaugh, J., Labrum, M. J., & Baker, R. S. (2020). Predicting math identity through language and click-stream patterns in a blended learning mathematics program for elementary students. *Journal of Learning Analytics*, *7*(1), 19-37.

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of Game-Based Learning. *Educational Psychologist, 50*(4), 258–283. https://doi.org/10.1080/00461520.2015.1122533

Shute, V. J., Almond, R. G., & Rahimi, S. (2019). *Physics Playground (Version 1.3)* [Computer software]. Tallahassee, FL. https://pluto.coe.fsu.edu/ppteam/pp-links/.

Shute, V. J., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C.-P., & Sun, C. (2020). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity, and learning supports in educational games. *Journal of Computer-Assisted Learning*, 1–15. https://doi.org/10.1111/jcal.12473

Acknowledgments

This work was supported by the US National Science Foundation [Award Number #037988]. We also would like to acknowledge Russell Almond, Fengfeng Ke, Zhichun Liu, Jiawei Li, Ginny Smith, Chen Sun, Xi Lu, and Chi-Pu Dai for helping in different phases of this project.