This document is published in:

José A. Ruipérez-Valiente, Pedro J. Muñoz-Merino, Carlos Delgado Kloos. A Predictive Model of Learning Gains for a Video and Exercise Intensive Learning Environment. In 17th International Conference on Artificial Intelligence in Education, 2015.

The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-19773-9_110

A Predictive Model of Learning Gains for a Video and Exercise Intensive Learning Environment

José A. Ruipérez-Valiente[a,b], Pedro J. Muñoz-Merino[a], Carlos Delgado Kloos[a]

[a] Universidad Carlos III de Madrid, Avenida Universidad 30, 28911 Leganés (Madrid) Spain
[b] IMDEA Networks Institute, Av. del Mar Mediterráneo 22, 28918 Leganés (Madrid) Spain
{jruipere, pedmume, cdk}@it.uc3m.es

Abstract. This work approaches the prediction of learning gains in an environment with intensive use of exercises and videos, specifically using the Khan Academy platform. We propose a linear regression model which can explain 57.4% of the learning gains variability, with the use of four variables obtained from the low level data generated by the students. We found that two of these variables are related to exercises (the proficient exercises and the average number of attempts in exercises), and one is related to both videos and exercises (the total time spent in both) related to exercises, whereas only one is related to videos.

Keywords: educational data mining, prediction, learning analytics

1 Introduction

There is a natural wish to be able to predict a future outcome. Prediction on education field has been extensively research over the years. The targeted objectives in education are diverse, for example to be able to predict the score of a future test [1, 2]. The increased use of Massive Open Online Courses (MOOCs) over the last few years provides of a perfect scenario to apply prediction techniques with large amounts of data. The high number of enrollments in these courses makes impossible to monitor each student separately, as an example the work by Brinton *et al.* [3] analyzes data from more than 100.000 students taking MOOCs in Coursera platform; therefore it is necessary the implementation of artificial intelligence tools to support and improve the learning process.

This work aims at proposing a predictive model of learning gains by using some variables which have been obtained through an experience using MOOC technology, specifically the Khan Academy platform. However the access was restricted to a predefined number of students in what is so called Small Private Online Courses (SPOCs). In this approach we have selected low level variables as predictors, which can be retrieved in a straightforward

way from the learning environment. We can find other works which use similar variables such as *avg_attempts* or *total_minutes* to predict students' test scores [1, 2]. In addition, other studies use different variables such as Pardos and Baker [4] where the predictor variables represent affective states.

2 Description of the Experience

The experience is framed in the 0-courses taken by freshmen students at Universidad Carlos III de Madrid (UC3M). A personalized Khan Academy instance with exercises and videos developed by the instructors of UC3M was provided. In this experience, the main educational resources were exercises and videos. For these experiences we have also enabled our learning analytics tool ALAS-KA [5], which implements many of the parameters that we use in this prediction model.

This research has been conducted in the chemistry and physics courses of 2014. Courses were composed of 51 exercises and 24 videos for chemistry, and 33 for both exercises and videos for physics. We have designed a pre-test and post-test from a pool of questions of equal hardness for both physics and chemistry. We define a student learning gain by obtaining the difference between post-test minus pre-test (LG = post – pre). We obtained only a total amount of valid samples of 44 students in physics and 25 in chemistry which were incorporated into the prediction model.

We have selected and retrieved a set of low level variables which are related to the learning process. The variables that we have considered are the *pre_test_score*, the *pre_test_time*, *correct_exercises* (percentage of correct exercises that the student tried to solve), *exercises_solved_once* (percentage of different exercises that were solved at least once), *proficient_exercises* (percentage of exercises in which the student has acquired a proficiency level), *avg_hints* (average number of hints in exercises), *avg_attempts* (average number of attempts in exercises), *avg_video_progress* (average progress by the student in all videos of the course), *videos_completed* (percentage of videos completed by the student), *total_time* (total time spent in both videos and exercises by the student), *exercise_time* and *video_time*.

3 Prediction model and discussion

After an exploratory analysis with our data and a review of the state of the art, we proceed to make the selection of variables and performed the linear regression analysis. We selected a hierarchical method with two entry steps and a total of four independent variables (introducing two of them in each step). The ANOVA test proved that both models are better than the baseline prediction of the learning gain; the F-value of the first model ($F = 30.5$, $p = 0.000$) is a bit higher than the second one ($F = 21.6$, $p = 0.000$) due to the insertion of more predictors.

We can check the model summary in table 1. The first model (with just two variables) has an R^2 of 0.481, while the second model (with four variables) rises to 0.574 after adding up the two new variables. Therefore, our second model is able to account a 57.4 % of the variation in the learning gains. In addition, the standard error of prediction is 15.1 points. Table 2 shows the report for the coefficients of the predictor variables in the two models; we can take a look at the standardized coefficients to have a feeling about the importance of each predictor in the model. In addition, equation 1 shows the prediction model formula. Next, we make an analysis of the different predictor variables:

LG = 25.489 - 0.604 * pre_test_score + 6.112 * avg_attempts + 0.017 * total_time + 0.084 * proficient_exercises (1)

- *pre_test_score*: this is the most important predictor in the model. The negative sign implies that the higher is the initial knowledge of students, the lower is going to be the increment in their knowledge. For every point in the pre-test, the predicted learning gain decreases 0.604 points.
- *avg_attempts*: the average number of attempts in exercises was also found to be an important predictor, whereas others like the average number of hints or time in exercises were not as important. For every unit that the average number of attempts increases, the predicted learning gain increases 4.093 points.
- *total_time*: the total time spent in videos and exercises (in minutes) is the second most important predictor of the model. For every additional minute, the predicted learning gain increases 0.017 points. This relationship makes sense, because if the student spends more time doing learning activities, it is more probable that the student learns more.
- *proficient_exercises*: the percentage of proficient exercises is the least important variable of the model, which might be quite surprising as it is related to how much progress the student did on exercises on the platform. However we should also take into account that it was the last variable entered in the model and that the total time spent in the platform might imply a better performance.

We have only used four independent variables, which is a prudent number considering the number of cases of our data sample. Three of the selected variables are related to exercises (*avg_attempts*, *total_time* and *proficient_exercises*) while one is related to videos (*total_time*). An important aspect is that measures related only to video progress (*avg_video_progress* and *videos_completed*) were not found as important as the ones related to exercises. However, we should state that progress in videos was also a useful predictor, but progressing on exercises variables had a more powerful impact in the model. It is also noteworthy to say that there are only three cases with a standardized residual above ±2, and none of them is over ±2.7, which means that there are not outliers. Thus the model is well fitted. The number of cases

in the data sample was too small to make a cross-validation. However, we can argue that all the assumptions (linearity, independence of variables and errors, homoscedasticity, multicollinearity, normally distributed errors) from the regression model were fulfilled, thus the model should generalize properly in experiences under a similar context and variables.

One of the issues from these results is that, while the pre-test variable was the most important predictor of the model, sometimes it is not feasible to have the initial knowledge of the students (via pre-test or from a different source). A future research question is if these results can be extrapolated to different platforms such as Open edX with different indicators and types of exercises. As part of future work, we would like to use new variables which provide higher level information such as student behaviors, students' efficiency or by the combination of different powerful predictors.

Model	*R*	*R Square*	*Std. Error of the Prediction*
1	0.693	0.481	16.42
2	0.758	0.574	15.1

Table 1. Model summary of the linear regression model.

Model	*Independent Variable*	*Un-std. Coeff.*		*Std. Coeff.*
		B	*Std. Error*	*Beta*
1	Constant	38.556	7.88	
	pre_test_score	- 0.601	0.84	- 0.655
	avg_attempts	4.093	3.149	0.119
2	Constant	25.489	8.071	
	pre_test_score	- 0.604	0.08	- 0.658
	avg_attempts	6.112	3.134	0.177
	total_time	0.017	0.011	0.202
	proficient_exercises	0.084	0.084	0.134

Table 2. Coefficients of the regression model.

Acknowledgements

This work has been supported by the "eMadrid" project (Regional Government of Madrid) under grant S2013/ICE-2715 and the EEE project (Spanish Ministry of Science and Innovation, "Plan Nacional de I+D+I) under grant TIN2011-28308-C03-01

References

1. Feng, M., Heffernan, N., Koedinger, K.: Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. Proceedings of 8th International Conference, Intelligent Tutoring Systems Jhongli, Taiwan. pp. 31–40. Springer-Verlag, Berlin (2006).
2. Feng, M., Beck, J., Heffernan, N., Koedinger, K.: Can an Intelligent Tutoring System Predict Math Proficiency as Well as a Standarized Test? In: Baker and Beck (eds.) Proceedings of the 1st International Conference on Educational Data Mining. pp. 107–116. Montreal (2008).
3. Brinton, C., Chiang, M., Jain, S., Lam, H., Liu, Z., Wong, F.: Learning about social learning in MOOCs: From statistical analysis to generative model. IEEE Transactions on Learning Technologies. 7(4), 346 - 359 (2014).
4. Pardos, Z., Baker, R.S.: Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. J. Learn. Anal. 1, 107–128 (2014).
5. Ruipérez-Valiente, J. A., Muñoz-Merino, P.J., Leony, D., Delgado Kloos, C.: ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform. Journal of Computers in Human Behavior. 47, 139 - 148 (2015).