Predicting Learners' Behaviours to Get it Wrong

Niels Heller, François Bry

Institute for Informatics, Ludwig-Maximilian University of Munich, Germany

Abstract. One of the most vexing aspects of tertiary education is the learning behaviour of many beginners: Late drop-outs after much time has already been invested in attending a course, incomplete homework even though completed homework is a sufficient condition for success at examinations, and misconceptions that are not overcome early enough. This article presents three predictors related to these learning-impairing behaviours that have been built from data collected with a learning platform and by examining homework assignments, and developed as Hidden Markov Model, by relying on Collaborative Filtering, and by using Multiple Linear Regression. The sensitivities and specificities of the first two predictors are above 70% and the R^2-error of the third predictor is about 20%. Considering the large numbers of unknown parameters like course-independent learning, this quality is satisfying. The predictors have been developed for fostering a better learning by raising the learners' consciousness of the deficiencies of their learning. In other words, the predictors aim at "getting it wrong". The article reports on the predictors and their evaluation.

1 Introduction

One of the most vexing aspects of tertiary education, especially in Science, Technology, Engineering, and Mathematics (STEM), is the learning behaviour of many beginners: Late drop-outs after much time has already been invested in attending a course, incomplete homework even though completing homework is known to be a sufficient condition for success at STEM examinations, and misconceptions or common fallacies related to a misunderstanding of mathematical and other abstractions that are not overcome early enough.

Late drop-outs, incomplete homework and misconceptions are arguably among the main reasons for failing at examinations. Reasons for these symptoms include the difficulty for beginners to adjust to a teaching style significantly different from that of secondary schools and the very limited individual feedback from experienced teachers resulting from high numbers of students per teacher (commonly 70 or more in computer science at German universities).

The research reported about in this article aims at compensating for the insufficient human coaching of STEM students for their better learning with algorithmically-generated individual feedback. Using predictors referring to the afore-mentioned three learning obstacles –late drop-outs, incomplete homework, and misconceptions– learners are made conscious of currently sub-optimal aspects of their learning and are motivated to improve it. This article reports on the first stage of this endeavour, the development and evaluation of three predictors of the afore-mentioned three learning obstacles. More precisely, the three predictors respectively forecast:

1. Skipping, that is, not fully participating or not participating at all in learning activities such as a weekly homework or a lecture
2. Examination fitness measured as a mark, that is, a percentage of the total score obtainable at an examination
3. Misconceptions related to a course's content

All three predictors have been developed using data on learners' behaviours collected from computer science courses given at Ludwig-Maximilian University of Munich. Some data has been collected through the teaching and learning platform Backstage [16, 3], other by examining homework assignments and examinations. The Skipping Predictor is a Hidden Markov Model [18], the Examination Fitness Predictor is a Multiple Linear Regression Model [4, chapter 6] and the Misconception Predictor relies on Collaborative Filtering [13]. The predictors' qualities are as follows:

1. Skipping is predicted with a sensitivity of 72.9% and a specificity of 84.7%.
2. Examination fitness is predicted with a R^2-error of 20.0% in one dataset and 19.8% in another.
3. Misconceptions are predicted with a sensitivity of 71.8% and a specificity of 80.7%

This article consists of 7 sections. Section 1 is this introduction. Section 2 presents related work. Each of sections 3, 4, and 5 describes one of the three predictors, reports on its quality, and on the individual feedback it gives to learners. Section 6 reports on perspectives for future work. Section 7 is the conclusion.

2 Related Work

This article is a contribution to "learning analytics", that is, "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [21]. Most learning analytics refer to Massive Open Online Courses (MOOCs). In contrast, the learning analytics presented in this article refer to formal education and presence courses.

Drop-out Prediction. There is no widely accepted definition of drop-out. Some authors define it as a discontinued participation in a formal course of study [14]. Other authors define it in terms of periods of inactivity during a course that can span from several weeks [11] to several years [22]. The Skipping Predictor presented below in Section 3 refers to skipping defined as not taking part in the next learning activity like a lecture or a homework assignment. Skipping is thus a narrow form of drop-out. Drop-out predictions often rely on measures of engagement in, or satisfaction with, a course [7, 9]. In contrast, the Skipping Predictor presented below is based on learners' gaps in knowledge. The drop-out predictor presented in [14] uses both, time invariant data gathered from registration forms and time dependent data gathered from multiple-choice tests and aims at good predictions already at the beginning of a course. Different methods have been used for drop-out prediction: Support Vector Machines [14], Neural Networks [14, 10], Decision Trees and Bayesian Classifiers [8].

Examination Performance Prediction. Both, time invariant data (such as grades in previously attended courses or demographic data) and time dependent data (such as engagement measures) have been used in predicting examination performances [6, 1, 12]. The predictors described in that articles use Neural Networks, Ordinal Logistic Regression, and Multiple Linear Regression, respectively. The Examination Fitness Predictor presented below in Section 4 is based on learners' participation in learning activities like lectures and homework assignments. Indeed, that predictor's aim is to suggest remedies in case insufficient examination performances are predicted. Basing predictions on parameters (like demographic data) learners cannot influence would be counterproductive and unethical. A correlation between emotional affects and examination performances has been observed by Pardos et al. [15]. Relying on experts for estimating learners' affects, these authors built a dataset and used it in training a machine learning method.

Learners' Misconceptions. The predictor of learners' misconceptions presented below in Section 5 is to the authors' knowledge the first of its kind. It is in the constructivist tradition of Posner et al. [17] who proposed the "conceptual change model" stating that learners have some, possibly erroneous, conceptions when they engage in a learning activity and that these conceptions change when certain conditions are met. That model is widely applied in science education [20] and is the basis of the Misconception Predictor presented below. That predictor also relates to research on provoking conceptual changes among learners [23] and on Radatz' cognitivist investigations of procedural errors in mathematical problem solving [19]. Procedural errors relate to misconceptions because "they result from non random applications of rules based on certain beliefs" [5, p. 33].

3 Skipping Predictor

The Skipping Predictor is based on a human labelling of homework submissions according to the following scheme:

- **SKIP**, for skipped, when a homework assignment is not delivered.
- **IK**, for insufficient knowledge, reflected by an incorrect use of symbols, statements like "I don't know how to solve this", or an answer not fitting a question, and requiring learning again part of the course material.
- **OE**, for other error, that is, errors not due to an insufficient knowledge.
- **NE**, for no errors, otherwise.

The homework submissions of 80 randomly selected students enrolled in an introductory course on computer science theory have been used in building the Skipping Predictor. The course had 11 weekly homework assignments. 80 label sequences of length 11 were thus generated. The labelling has been performed by the course's teaching team. For testing purposes, four members of the teaching team categorized independently from each other 30 of the $80 \times 11 = 880$ submissions yielding an inter rater reliability Fleiss-κ of 0.78. For evaluation purposes, one member of the teaching team categorized once again all 880 submissions.

The predictor is a Hidden Markov Model, that is, a dynamic state system consisting of "hidden states" of which one is currently active [18]. The active state changes at each step, that is, with each weekly homework, according to predetermined state transition probabilities. A state depends on all assignment submissions of each student so far. An "emission" (or "observation"), taken from a fixed set, is observed in each step. The probability of observing an emission is defined by the active state. The afore-mentioned set of labels –SKIP, IK, OE, and NE– was used as emission set. Two hidden states were used in the model. The model was trained using the Baum-Welch Algorithm [2]. The evaluation was a 10-fold cross validation.

The predictors quality is, as usual, estimated by its sensitivity (or recall), that is, the proportion of predicted skips that have taken place, and by its specificity" (or true negative rate), that is, is the proportion of non-skips that have been predicted. The Skipping Predictor has a sensitivity of 72.9% and a specificity of 84.7%.

These results are satisfying, yet higher inter rater reliability of the labels and specificity would be preferable. Indeed, it is preferable to nudge against skipping as few learners as possible that are not at risk of skipping. We expect the labellers' competence to improve over time what should result in a higher inter rater reliability and a higher specificity.

4 Examination Fitness Predictor

The Examination Fitness Predictor is based on the numbers of SKIP, IK, OE, and NE labels described in the previous section 3 assigned to the homework submissions of each student. The predictor applies a Multiple Linear Regression Model, that is, it expresses a dependent numerical variable as a linear combination of independent numerical variables. The predictor's dependent variable is the examination fitness expressed as a mark, that is, a percentage of the total score obtainable at the examination. Its independent variables are the numbers of each label SKIP, IK, OE, and NE assigned to the homework submissions of each student.

Two Examination Fitness Predictors have been built from two datasets. The first dataset, the "homework dataset", consists in the label numbers of the dataset described above in section 3. The second dataset, the "weekly learning assessment dataset", consists in the label numbers of quiz answers collected by the system Backstage [16, 3] during an introductory course on functional programming. Using Backstage, a multiple-choice quiz, each possible answer of which had been labelled SKIP, IK, OE, or NE, was run at the beginning of each weekly lecture so as to assess how well the content of the last week's lecture had been learned. The second dataset is needed for the following reasons: It allows an evaluation on data that do not require any human labelling and on data referring to a different educational setting: Answering quizzes on last week lecture and homework assignment on the current week lecture, respectively. Interestingly, with these datasets different independent variables, including in both cases the variable IK (insufficient knowledge), have been identified as impacting on examination fitness:

– With the homework dataset, SKIP (missed learning activity) and IK (insufficient knowledge) impact on examination fitness with a significance level of 0.05%.
– With the weekly learning assessment dataset, OE (other error) and IK (insufficient knowledge) impact on examination fitness with a significance level of 0.05%.

The Examination Fitness Predictor's quality, expressed as usual for a Multiple Linear Regression Model by the coefficient of determination (R^2), is as follows:

– Homework dataset with the independent variables SKIP and IK: $R^2 = 20.0\%$
– Weekly learning assessment with the independent variables OE and IK: $R^2 = 19.8\%$

Examination Fitness Predictors could yield better predictions if, in addition to the afore-mentioned variables, they would refer as well to measures of activity and to demographics. We rejected such an improvement for two reasons. Firstly, the resulting predictions could wrongly suggest to students that more activity, whatever its nature, could positively impact on examination fitness. Secondly, demographics cannot be influenced by students. Since the goal of our predictors is to nudge students to a better learning, the predictors' quality must be subordinated to the impact of their predictions on the students' behaviours.

The forecasts of the Skipping and Evaluation Fitness Predictors allows for the following intervention. The Skipping Predictor recognizes students at risk of skipping. The Evaluation Fitness Predictor forecasts that skipping a homework assignment reduces examination fitness on average by $2,5\%$, as the evaluation has shown. Such an automatically generated feedback keeps many, if not all, students "on tracks", by nudging them not to skip their next homework assignment. A description of this nudging and its evaluation are out of the scope of this article.

5 Misconception Predictor

The Misconception Predictor relies, like the Skipping Predictor, on a human labelling of the students' weekly homework submissions. In contrast to the Skipping Predictor, however, no pre-defined labels were used. Instead, the labellers introduced new labels as they needed them. The labelling was performed on the project platform of our Backstage 2 system what ensured the sharing of misconception labels between the labellers. The homework submissions of 80 students enrolled in an introductory course on the theory of computer science have been labelled by the teaching team in two steps: Firstly, a labeller read a large number of submissions for a specific assignment so as to identify and label common misconceptions. Secondly, all submissions were labelled according to the identified misconceptions. This approach is the standard correction routine of teaching teams. In order to test whether the generated labels were reliable, two measures were computed. Firstly, two labellers independently examined different sets of submissions for 4 assignments. Both labellers recorded those misconceptions they frequently encountered. Even though they were not told how many frequent misconceptions they should record, they both recorded as frequent 2 to 3 misconceptions. For 3 of the 4 examined assignments, the labellers fully agreed. For the remaining assignment, the labellers agreed on two misconceptions but they reported a different third one. Secondly, 4 further labellers where given descriptions of the frequent misconceptions formerly recorded and independently labelled 20 other submissions. On average and for all misconceptions, the inter rater reliability was Fleiss-$\kappa = 0.70$.

The Misconception Predictor relies on collaborative filtering [13], following a simple intuition: If two learners had similar (specific) misconceptions in the past, they are likely to have similar misconceptions in the future.

10-fold cross validation yielded the following quality measures:

- Sensitivity 71.8%
- Specificity 80.7%

First investigations with data collected from different lectures point to the effectiveness of the approach to build Misconception Predictors. Predictions on misconceptions are highly beneficial for keeping learners "on tracks" by warning them against "shallow learning". This nudging is out of the scope of this article.

6 Perspectives for Future Work

The predictors presented above are based on automatically gathered data using multiple-choice quizzes, and data collected by a human labelling that exhibits a high inter rater reliability. We expect the teaching teams' competence to improve over time what should result in a higher inter rater reliability and a higher specificity. Peer review, or peer assessment, that is, the enrolment of students to providing feedback, in place of or in addition to the teacher's feedback, to their fellow students, could also explicitly or implicitly contribute to a better labelling, especially of misconceptions. Adopting peer review in this manner is likely to be well accepted since students see three main benefits in peer reviews: the chances "to compare different approaches", to "compare standard of work," and the "exchange of information and ideas" [24, p. 52].

An automatic labelling of student behaviour can be easily realized for tasks such as multiple-choice quizzes or programming (unit tests can be used to asses programs). The labelling of student behaviour on other tasks might require advanced methods such as machine.

The approach described above is a proof-of-concept. For deploying its full potential, it has to be applied with datasets referring to many more courses and many more course venues. We expect that a clustering of courses will result in improved predictors for the course clusters. Indeed, the educational context is likely to matter: Is for example the students' skipping behaviours in a Calculus and Linear Algebra courses similar? In programming courses?

7 Conclusion

This article has presented three predictors, the Skipping, Examination Fitness, and Misconception Predictor, related to learning-impairing behaviours, skipping (a form of drop-out), improper homework, and misconceptions. The predictors have been built from data collected from courses in computer science using the system Backstage [16, 3] and by examining homework assignments. One predictor is a Hidden Markov Model, a second predictor relies on Collaborative Filtering, and the third predictor uses Multiple Linear Regression. The sensitivities and specificities of the first two predictors are above 70% and the R^2-error of the third predictor is about 20%. Considering the large numbers of unknown parameters like course-independent learning, this quality is satisfying.

The predictors have been developed for giving an algorithmically generated feedback to students, so as to make them conscious of deficiencies in their learning and to nudge them to improve it. Thus, the predictors have been developed for "getting it wrong" because warned students better their learning and, in doing so, invalidate the predictors' forecasts. Hints at this use of the predictors, an issue out of the scope of this article, have been given. The authors expect that extensive evaluations will point to the effectiveness of nudging students towards better learning with the predictions described in this article.

References

1. Abdous, M., Wu, H., Yen, C.J.: Using data mining for predicting relationships between online question theme and final grade. Journal of Educational Technology & Society **15**(3), 77 (2012)
2. Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics **41**(1), 164–171 (1970)
3. Bry, F., Pohl, A.Y.S.: Large class teaching with Backstage. Journal of Applied Research in Higher Education **9**(1), 105–128 (2017)
4. Christensen, R.: Plane answers to complex questions: the theory of linear models. Springer Science & Business Media (2011)
5. Confrey, J.: Chapter 1: A review of the research on student conceptions in mathematics, science, and programming. Review of Research in Education **16**(1), 3–56 (1990)

6. Cripps, A.: Using artificial neural nets to predict academic performance. In: Proceedings of the 1996 ACM Symposium on Applied Computing, pp. 33–37. ACM (1996)
7. Dejaeger, K., Goethals, F., Giangreco, A., Mola, L., Baesens, B.: Gaining insight into student satisfaction using comprehensible data mining techniques. European Journal of Operational Research **218**(2), 548–562 (2012)
8. Dekker, G., Pechenizkiy, M., Vleeshouwers, J.: Predicting students drop out: A case study. In: Educational Data Mining 2009 (2009)
9. Giesbers, B., Rienties, B., Tempelaar, D., Gijselaers, W.: Investigating the relations between motivation, tool use, participation, and performance in an e-learning course using web-videoconferencing. Computers in Human Behavior **29**(1), 285–292 (2013)
10. Guo, W.W.: Incorporating statistical and neural network approaches for student course satisfaction analysis and prediction. Expert Systems with Applications **37**(4), 3358–3365 (2010)
11. Halawa, S., Greene, D., Mitchell, J.: Dropout prediction in MOOCs using learner activity features. eLearning Papers **37** (2014)
12. Huang, S., Fang, N.: Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers & Education **61**, 133–145 (2013)
13. Linden, G., Smith, B., York, J.: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing **7**(1), 76–80 (2003)
14. Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., Loumos, V.: Dropout prediction in e-learning courses through the combination of machine learning techniques. Computers and Education **53**(3), 950–965 (2009)
15. Pardos, Z.A., Baker, R.S., San Pedro, M.O., Gowda, S.M., Gowda, S.M.: Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 117–124. ACM (2013)
16. Pohl, A.: Fostering awareness and collaboration in large-class lectures. Doctoral thesis, Ludwig-Maximilian University of Munich, Germany (2015)
17. Posner, G.J., Strike, K.A., Hewson, P.W., Gertzog, W.A.: Accommodation of a scientific conception: Toward a theory of conceptual change. Science Education **66**(2), 211–227 (1982)
18. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)
19. Radatz, H.: Error analysis in mathematics education. Journal for Research in Mathematics Education pp. 163–172 (1979)
20. Read, J.R.: Childrens misconceptions and conceptual change in science education (2004). URL http://asell.org/global/docs/conceptual_change_paper.pdf
21. Siemens, G. (ed.): 1st international conference on learning analytics and knowledge 2011 (2010)
22. Tan, M., Shao, P.: Prediction of student dropout in e-learning program through the use of machine learning method. iJet **10**(1), 11–17 (2015)
23. Tippett, C.D.: Refutation text in science education: A review of two decades of research. International Journal of Science and Mathematics Education **8**(6), 951–970 (2010)
24. Williams, E.: Student attitudes towards approaches to learning and assessment. Assessment and Evaluation in Higher Education **17**(1), 45–58 (1992)