Developing predictive models for early detection of at-risk students on distance learning modules

Annika Wolff Zdenek Zdrahal Drahomira Herrmannova Jakub Kuzilek Martin Hlosta
Knowledge Media Institute, The Open University
Walton Hall
Milton Keynes, MK7 6AA
+44(0)1908 659462, 654512, 652477, 659109, 653800
a.l.wolff;z.zdrahal;drahomira.herrmannova;jakub.kuzilek;martin.hlosta{@open.ac.uk}

ABSTRACT

Not all students who fail or drop out would have done so if they had been offered help at the right time. This is particularly true on distance learning modules where there is no direct tutor/student contact, but where it has been shown that making contact at the right time can improve a student's chances. This paper explores the latest work conducted at the Open University, one of Europe's largest distance learning institutions, to identify when is the optimum time to make student interventions and to develop models to identify the at-risk students in this time frame. This work in progress is taking real-time data and feeding it back to module teams as the module is running. Module teams will be indicating which of the predicted at-risk students have received an intervention, and the nature of the intervention.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining; D.4.8 [**Performance**]: Modelling and prediction

General Terms

Algorithms, Design, Experimentation, Human Factors, Theory.

Keywords

predictive models, machine learning, student data, Bayesian models, distance learning

1. INTRODUCTION

Predictive modelling techniques can be applied to student data to identify students who are at risk of failing or withdrawing from a module. Tutors or module teams can use this information to aid their decision-making about whom they should contact to offer help, leading to better strategic use of resources and improved retention. For example, the Course Signals system has been successfully in place at Purdue University for some time, providing feedback to students based on predictions from multiple data sources (Arnold and Pistilli, 2012). The Open University (OU) is one of the largest distance learning institutions in Europe. OU modules are making increasing use of the Virtual Learning Environment, Moodle, to supply learning materials, instead of the previous paper materials supplied in the post.

This paper explores the latest work at the Open University using data from VLE, combined with demographic data to predict student failure or dropout. This ongoing work is already providing real-time information to module teams and will be fully evaluated later in the year. The methods investigate the role of VLE activity compared with demographic data and attempt to make predictions of a student before they submit their first assessment. This first assessment has been found to be a very good predictor of a students final outcome on a module.

This work is the culmination of a number of previous projects to investigate the potential for different methods to produce accurate predictions. We will first describe briefly some of the previous work at the OU before examining the current methods, preliminary feedback of these and plans for future evaluation.

2. Previous work with OU data

Decision trees have proved a fairly popular method for exploring the potential for building predictive models from student data (see Baradawaj and Pal, 2011; Pandey and Sharma, 2013; Kabra and Bichkat, 2011). Initial work with OU data focused on using decision tress to predict student outcome from VLE data combined with assessment scores. Each OU module evaluates students periodically with a Tutor Marked Assessment (TMA). The exact number may vary from module to module, but seven TMA's is quite typical. Three modules, each with fairly typical VLE usage and a large student cohort (between 1200 and 4400 students registered), were chosen for building and testing the models. The main findings from this were that decision trees were fairly good at predicting both a drop in performance in a subsequent assessment and in predicting the final outcome of the module. Prediction was overall better when combining VLE and TMA data. This preliminary work also suggested that the absolute amount of clicking within the VLE was not directly correlated with outcome, students could click a lot but still fail or not click at all and still pass. However, reduction in clicking was a warning sign.

The models were developed and tested on single presentations of the three modules, then they were tested on a future presentation of the same module. Finally, they were tested on each other (in

other words, developed on one module and applied to another). As expected, accuracy was reduced in the latter two cases, but interestingly not as much as might have ben expected. A brief investigation into including demographic data revealed that prediction could be improved with this data source. This work is described in detail in Wolff et al. (2013a).

The next phase of work investigated more fully the potential for using demographic data and focused on Bayesian models for prediction, which were compared with more simple methods of linear and logistic regression and weighted score. The key findings were that a) including VLE data improved the accuracy of predictions compared to using demographic data alone b) there was little real difference between the different methods evaluated - accuracy increased as the module progressed. However, the majority of dropout occurs early in the module (Wolff et al. 2013b).

Some focused investigation into the role of the first TMA in predicting the final outcome, found that failing the first assessment had a significant negative impact. Therefore, the key to improving retention is in identifying those students who are at risk of either submitting but failing, or not submitting this first assessment. This is described in more detail in the next section, where the overall problem is specified.

3. Problem Specification

For identifying students at risk we can use knowledge about students' behavior and performance in the current presentation, their demographic data and data about the module and performance of others students in previous presentations. In this task we do not consider students' overall learning objectives, nor their previous or current performance in another modules. This is diagrammatically shown in Figure 1. A_1-A_n indicate the time at which a module assessment is due. Vle_1-Vle_n are the VLE clicks in the periods between either the start of the module and first assessment, or else in between assessments.

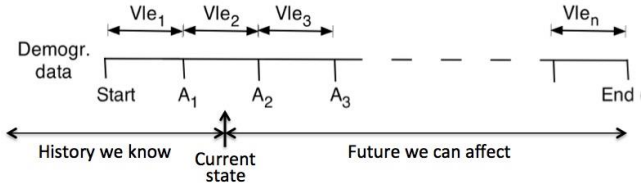

Figure 1. Prediction problem

Given demographic data, the results of TMAs so far and VLE activities, the goal is to identify as early as possible the students who are at risk and for whom the intervention is meaningful. By meaningful intervention, we mean that the student can be helped to pass the module and the overall cost of interventions is affordable. The reasoning about the future behavior of the student is based on experience with students with similar characteristics in previous presentations of same module.

Our analysis indicates that VLE data is more important than demographic characteristics. Moreover, the performance at the early stages of the module presentation is a very good predictor of final success or failure. In the analysed modules, the students who fail or do not submit the first TMA have high probability of overall failure. For this reason it is crucial to concentrate the effort to identify at risk students before the TMA1 deadline. This is indicated in Figure 2.

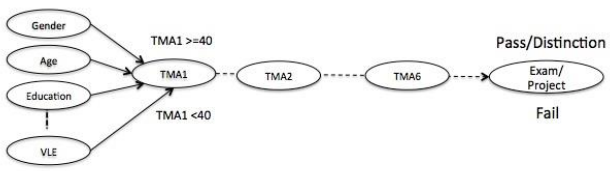

Figure 2. TMA1 is a strong predictor of the final result

The VLE opens two to three weeks prior to the start of the module presentation so that students can smoothly engage early in a number of module related activities. In order to achieve early predictions for TMA1 we start analysing records from the very opening of the VLE, i.e. well before the presentation start. VLE activities can be classified into a number of actions and activity types depending on what is the student trying to do. Out of many, we have identified four *activity types* that provide useful information for prediction. They are denoted as follows:

- **R**esource contains books and other educational materials for the students
- **F**orum is a web site where students communicate with their tutors and with each other
- **S**ubpage is the means of navigation in the VLE structure
- **O**U Content refers to the specification of TMAs and the guidelines for their elaboration.

Our predictive modeling algorithms use, for each student, weekly aggregates of all four activity types and all their combination. Therefore, for each student and each week we get a 16 dimensional vector (N, R, F, S, O, RF, RS, …, RFSO) where N means "No VLE activity". Some algorithms use numeric values describing the number of accesses of particular activity type, others use mutually exclusive Boolean values representing the fact that the student engaged in the particular combination of activity types.

4. Methods for early detection of failure

Predictions of at risk students is calculated and updated every week starting from the opening of the VLE. The prediction combines the results of four machine learning algorithms:

1. **k Nearest Neighbours** (k-NN) makes use of weekly aggregates represented as 16-dimensional numerical activity type vectors compared with legacy data of previous presentations.
2. **k Nearest Neighbours** (k-NN) is based on a similar approach but uses only demographic data. Since demographic data has typically nominal values, an important part of the algorithms was how to define distance between two demographic sets.
3. **Classification and Regression Tree** (CART) is calculated from VLE data and TMA1 of previous presentations and then used for the classification of current students.
4. **Bayes network** combines both demographic and VLE data. Chi-square tests showed that a statistically independent subset of demographic data exist. For a smaller number of demographic variables a full Bayes network has been constructed. For the complete set, we implemented naïve Bayes.

The results of these methods are combined by majority voting.

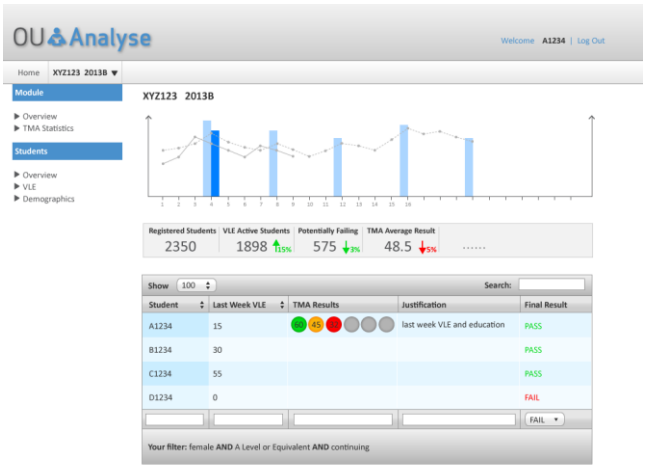

Figure 3a. Mockup module dashboard

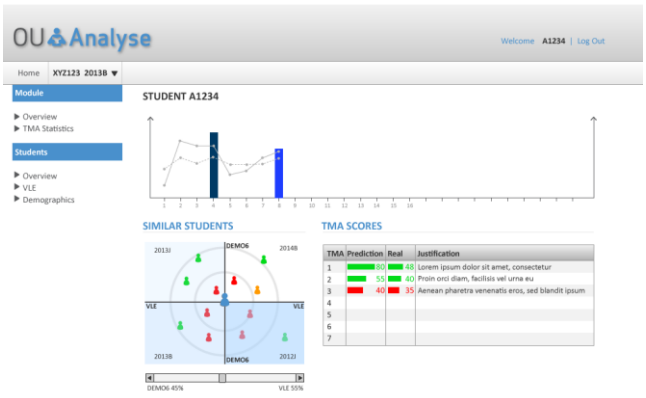

Figure 3b. Mockup dashboard describing an individual student

The mockups of the dashboards for presenting the results are shown in Figures 3a and 3b. Figure 3a demonstrates a view across students of a module. The upper graph presents an overview of VLE activities. The lower table organizes students according to their predicted outcome at the current point in the module, including an explanation for the prediction. Figure 3b shows the view of an individual student.

The detail of the interface that allows the tutor to change the balance of predictions based on demographic and VLE data is shown in Figure 4.

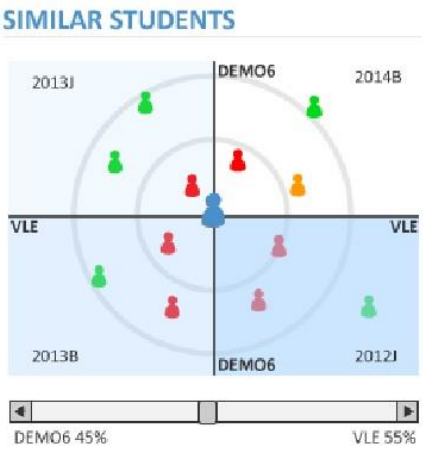

Figure 4. 3-nearest neighbours based on demographic and VLE distances

The icon representing the evaluated student is in the centre. The upper right quadrant shows the three nearest neighbours in the current presentation. The nearest neighbours of three previous presentations are organised anticlockwise. In the quadrants representing the previous presentations, the red icons indicate that the student failed, the green ones indicate a pass. In the current presentation the icons show predicted outcome. The amber icon show the borderline cases. The default split is calculated by the algorithms, but the tutors can express their experience by moving the slider.

The list of students identified as at risk is passed to the module team for possible interventions. Currently, the data is passed in a spreadsheet, whilst the dashboard mockups are being also evaluated with module teams and will be completed and integrated with models and data when the design is finalized. The spreadsheet rank orders the students on order of their weighted risk score. An explanation for the prediction of each of the models is given. The first two explanations point to the nearest neighbours from the previous presentations (first the closest by their VLE activity and secondly the closest by their demographic data). Next, the prediction according to the decision tree is explained in terms of the applied rule, which may combine the students level of VLE activity with some demographic attributes (these are the normal demographic attributes that are collected about students, e.g. age, previous academic background, etc.). Finally, the prediction of the Bayes classifier is presented along with the explanation similar to the decision tree, combining VLE with demographic information. In some cases, the predictions from the four models do not match.

5. Evaluation

Evaluation of the latest methods will occur when the module has completed. Regardless of the predictive methods being used, there is a general prediction by module teams that retention will improve in this presentation due to other factors, such as improved module design and also changes to student funding and the financial commitments that students are now making. This will clearly impact on the ability to draw any firm conclusions about what to attribute improved retention to, should that turn out to be the case. However, it is still worth looking at the overall retention compared to previous modules. It is also possible to use qualitative methods, such as looking at the student feedback, or speaking with the module tutors and module teams. In addition, it

is possible to make a hypothesis about the accuracy of the methods where interventions have been made. If interventions are having an effect, then this should reduce the accuracy of the predictions. Specifically, it should be the case that predictions made for a student prior to an intervention being made will give a false positive result for failure. The precision and recall of the methods on this module at this point in time can be compared to methods applied to other modules at the same point in time, to test for significant differences.

The first set of predicted outcome for TMA 1 has been provided to one of the pilot module teams and action will be taken in the very near future. While it is not possible to know yet what the final evaluation will show, the module team, as well as wider support networks for OU students, have been looking at the initial outputs and feel very positive about the potential for the technology to integrate into wider OU practice and provide an important source of information, both for strategically targeting support to students when they need it, but also for improving advice given to students as they begin their studies.

6. Conclusion

Where previous work has demonstrated that it is possible to accurately identify at risk students throughout a module presentation, this latest work focuses specifically on increasing accuracy for early detection. Most students who fail get into difficulties very early on, so this is the critical point at which to make an intervention. Predictions are made with reference to a students nearest neighbor, based firstly on demographic data and secondly on VLE data, allowing the two data sources to be balanced against each other and to better understand, over time, the role of each. In addition, CART and Bayes models are applied to the combined VLE and demographic data. Predictions from the four models are weighed against each other to produce a list of students ranked in order of risk. Currently, this is provided in a spreadsheet to module teams, along with explanations from each of the models. Dashboards are being constructed to visualize this data. The feedback from the first set of output data has been very positive. A full evaluation will occur later in the year when the module is complete.

7. REFERENCES

[1] Arnold, K.E., Pistilli, M.D. 2012. Course Signals at Purdue: Using Learning Analytics to increase student success. In: Learning Analytics and Knowledge, 29 April – 2 May, Vancouver, Canada

[2] Baradwaj, B., Pal, S. 2011. Mining Educational Data to Analyze Student's Performance, International. Journal of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63-69

[3] Pandey, M., Sharma, V.K. 2013. A Decision Tree Algorithm Pertaining to the Student Performance. Analysis and Prediction. International Journal of Computer Applications 61(13): 1-5, New York, USA

[4] Kabra, R. R. and Bichkar, R.S. 2011. Performance Prediction of Engineering Students using Decision Trees. International Journal of Computer Applications 36(11): 8-12, New York, USA

[5] Wolff, A., Zdrahal, Z., Nikolov, A., Pantucek, M. 2013a. Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In: *Third Conference on Learning Analytics and Knowledge (LAK 2013)*, 8-12 April 2013, Leuven, Belgium

[6] Wolff, A., Zdrahal, Z., Herrmannová, D. and Knoth, P. 2013b. Predicting Student Performance from Combined Data Sources, in eds. Alejandro Peña-Ayala, Educational Data Mining: Applications and Trends, 524, Springer