EVALUATING WEEKLY PREDICTIONS OF AT-RISK STUDENTS AT THE OPEN UNIVERSITY: RESULTS AND ISSUES

Drahomira Herrmannova, Martin Hlosta, The Open University, United Kingdom, Jakub Kuzilek, Zdenek Zdrahal, The Open University, United Kingdom, and CIIRC, Czech Technical University, Czech Republic

Introduction

Improving student retention rates is a critical task not only for traditional universities but particularly in distance learning courses, which are in recent years rapidly gaining in popularity. Early indications of potential student failure enable the tutor to provide the student with appropriate assistance, which might improve the student's chances of passing the course. Collated results for a course cohort can also assist course teams to identify problem areas in the educational materials and make improvements for future course presentations.

Recent work at the Open University (OU) has focused on improving student retention by predicting which students are at risk of failing. In this paper we present the models implemented at the OU, evaluate these models on a selected course and discuss the issues of creating the predictive models based on historical data, particularly mapping the content of the current presentation to the previous one. These models were initially tested on two courses and later extended to ten courses.

Problem description

The OU is one of the largest distance learning institutions in Europe, providing several hundred courses to more than 200,000 students. The students primarily receive their study materials online through a Virtual Learning Environment (VLE). Within each course students are split into study groups, typically based on their geographical location. Each study group is supported by a tutor who provides the students with advice, answers their questions, and marks their assessments and exams. In order to support the failing students, the university carries out various types of interventions, involving email, phone calls and face-to-face contact. However, as the resources (staff, time, etc.) available for interventions are limited and the number of students studying a course can reach several thousand, the interventions have to be carefully planned. The key issue is to identify students at risk of failing the course, but who might with additional support still succeed.

Machine learning methods are commonly used to build models for predicting student performance (Romero et al., 2013; Wolff et al., 2014; Huang & Fang, 2013). A typical

approach is to utilise historical data, for example data from the previous presentation of the course, to develop the predictive models, which can then be applied in the current presentation (Wolff et al., 2014; Huang & Fang, 2013). However, this approach is not always as straightforward as it sounds. In cases where the course structure has changed since the previous presentation (e.g. if the number of assessments or their cut-off dates changed), it may be difficult to correctly map the current presentation to the previous one. This might in turn decrease the accuracy of the predictive models.

Related work

A valuable predictor of student success is the results of the student's previous study (Huang & Fang, 2013). However, data from previously studied courses is not available for students who are attending their first courses. These courses are critical, as they determine whether the students will continue their studies. Moreover, retention in these courses is typically low (Wolff et al., 2014). Wladis et al. (2014) define predictive factors using demographic data for estimating student outcomes in online courses. Romero et al. (2013) summarise statistics about the student activity is used to predict student success.

The work presented here builds on the previous research at the OU (Wolff et al., 2014; Wolff & Zdrahal, 2012; Wolff, Zdrahal, Herrmannova & Knoth, 2013; Wolff, Zdrahal, Nikolov & Pantucek, 2013; Hlosta et al., 2014). Initial experiments were undertaken using student activity data from the VLE and their assessment scores (Wolff & Zdrahal, 2012; Wolff, Zdrahal, Nikolov & Pantucek, 2013). The solution presented in this paper combines both the students' demographic data and their activity in the VLE (combined data sources). The prediction is based on votes from four machine learning models. One of the key findings of our previous work was the importance of identifying failing students prior to the first assessment of each course (Wolff et al., 2014). We noticed that students who don't succeed in the first assessment are very likely to fail the whole course.

Data specification

Two types of data are collected: student demographic data and student VLE activity data, which represent individual actions (e.g. participation in forums, learning resources accessed, etc.). Data from the VLE are collected daily. At the OU, the course study plan is structured into weeks (every week, the students are expected to engage with new study materials and tasks). To match the data with the study plan we aggregate the student's activities into weekly summaries. Moreover, as the results of past assessments in the course become available, these are included in the models. Figure 1 shows the OU Analyse Dashboard, which depicts the average online activity during one course together with the average results of the assessment.
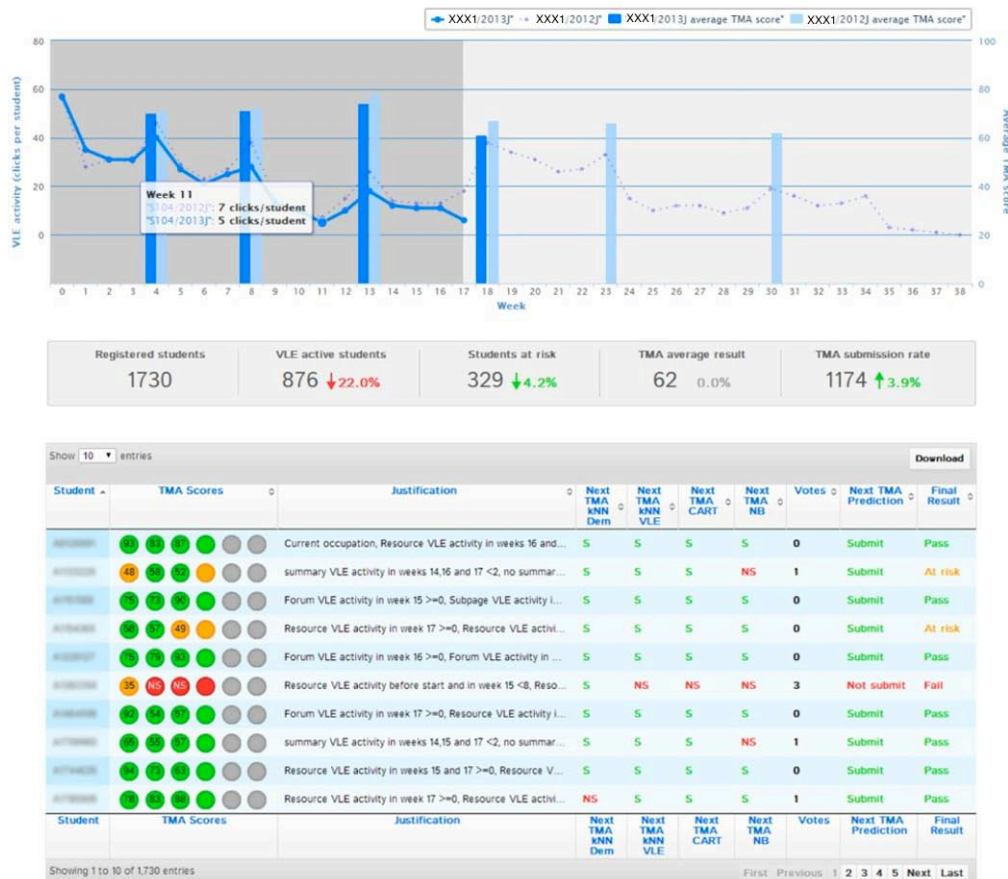
Figure 1. OU Analyse Dashboard showing average daily VLE activity for all weeks a course, course trends and individual student predictions. The bar chart shows the average students' results of the assessments.

Predictive models

Four models are used to predict whether the student will submit the next assessment: Classification and Regression Tree (CART); probabilistic Naïve Bayes model and k -Nearest Neighbours (k-NN), which is run separately using demographic and VLE data. The final verdict is achieved by voting, each model having the same weight. The student is predicted not to submit the assessment or fail if more than two models vote so.

Each model has been selected for its specific properties:

- The Naïve Bayes model makes decisions that are the most probable, making the model average error as small as possible (Duda & Hart, 1973).
- k-NN enables easy identification of the most similar students from previous presentations based on either VLE activity or on demographics. The error of k-NN is at most twice that of the Naïve Bayes model (ibid.).
- The CART method produces an easy to interpret set of rules represented as a decision tree. Moreover, CART works with both numerical and categorical attributes.

These models have been used as the basis for weekly at-risk student predictions for two courses in the summer semester 2014 and ten courses in the winter semester 2014. The predictions are sent every week to the course teams. In addition, all four models were migrated into the OU Analyse dashboard (analyse.kmi.open.ac.uk). The current set of models can be extended by adding new predictive models. The models can also be weighted by their varying importance during the voting process.

Model evaluation

The following section presents the performance evaluation of the four discussed models. The evaluation was done using the current presentation of one of the 10 courses, which we are now analysing. In each week, the goal was to predict if the student will or will not submit the next assessment. At the moment, the selected course has three completed assessments (denoted as A_1, A_2 and A_3), the last assessment cut-off date is in the 14th week of the course. We used F-measure as the metric for model performance comparison, which is a harmonic mean of the Precision and Recall measures. These are commonly used to compare machine learning algorithm performance. The use of a single measure allows us to easily compare the models.

The results of the evaluation are depicted in Figure 2. The three bold vertical lines represent the cut-off dates of the assessments A_1, A_2 and A_3 in weeks 4, 8 and 14. The F-measure for the given week relates to the prediction of the result of the following assessment. There are two interesting facts that are worth pointing out.

We can observe steep growth of the F-measure in the first week after both A_1 and A_2 for the demographic k-NN model. In addition to previous demographic features, the data in this week contains information about the results from the previous assessment. In contrast with the demographic attributes, the assessment result is highly discriminative. This can be seen in the first week after each assessment cut-off date when the value of the F-measure of the demographic k-NN rapidly increases.
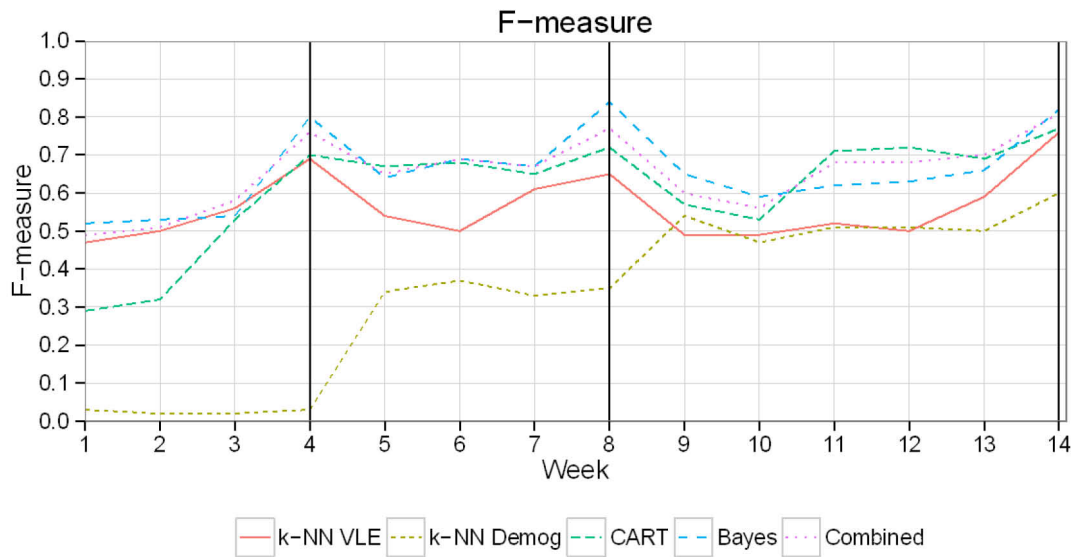
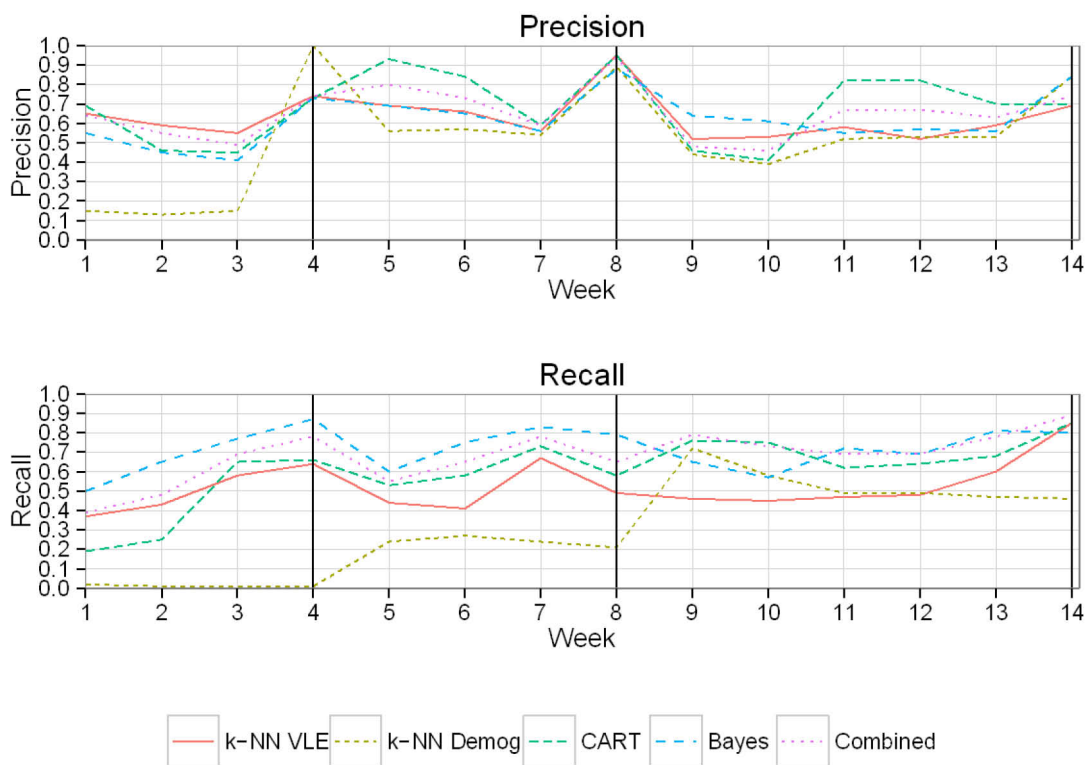Figure 2. F-measure for the first 14 weeks and 3 assessments of the selected course

Figure 3. Precision and Recall for the first 14 weeks and 3 assessments of the selected course

Furthermore, the importance of the VLE features (k-NN using VLE data) increases when approaching the cut-off date of the A_1 and decreases in the first week after the assessment. We can conclude that for predicting whether the student will submit the next assessment, it is more important what he/she does right before the assessment rather than at the beginning of

the assessment period. For example, after the assessment the students could take a short rest before focusing on the next one.

Issues

Although the performance of the models expressed by the F-measure looks promising, a more detailed inspection revealed some issues, that are visible after decomposing the F-measure into its components, Precision and Recall, as shown in Figure 3. Although the F-measure increases before the cut-off date as expected, Precision and Recall do not comply with the expected trend (see the drop in weeks 7-8 in Figure 3). This effect is caused by the fact that in the previous presentation (on which the models are trained) the cut-off date for the second assessment was one week earlier than in the current presentation.

Learning predictive models, assessments and prediction errors

We will refer to the previous presentation used for building the models as *model presentation* and to the current presentation in which students' results are predicted as *predicted presentation*. We assume that the structure of the course and the content of the study materials and assessments don't change between presentations, therefore, we can use the previous presentation for training the models. Each presentation has its start date, which can be any day of the calendar week. *Course weeks* are counted from the start date by repeatedly adding 7 days. Presentations typically have around 40 course weeks. Each course week can be associated with specific resources and tasks.

When trying to map the weeks between the model and the prediction presentations, three situations might occur. The ideal situation happens when the cut-off days for assessments coincide with the first day of some course week, and the number of weeks between assessments are the same for the model and predicted presentation, see Figure 4. In this case the predictions that take into account only student activities are not affected by errors due to data mismatch.
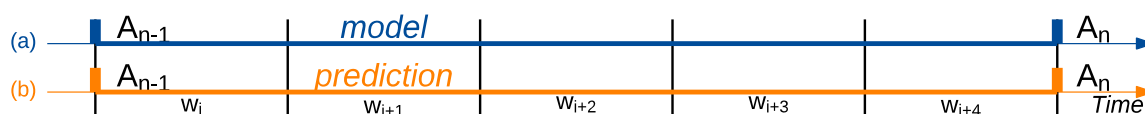

Figure 4. Ideal synchronisation between model and predicted presentations

Unfortunately, the reality is not always that simple. Very often the intervals between the corresponding assessments are not the same. This leads to a mismatch between planned activities in the corresponding weeks of the model and the predicted presentation, which may affect the precision and recall of predictions. If the assessment dates of the model and the predicted presentation differ, we may synchronise the presentations either at the beginning (*forward synchronisation*) or at the end (*backward synchronisation*) of the assessment period.

The models used to generate F-measure, Precision and Recall in Figures 2 and 3 use forward synchronisation (Figure 5), where the model and the predicted presentations are synchronised at the beginning of the assessment period. The period between assessments A_{n-1} and A_n in the

model presentation (Figure 5 (a)) can be shorter (Figure 5 (b)) or longer (Figure 5 (c)) than in the predicted presentation. In this case the data from the last week w_{i+3} before the cut-off date either has to be duplicated (Figure 5 (a) vs. 5 (b)) or merged with week w_{i+2} (Figure 5(a) vs. 5 (c)). In either case we can expect an increase in the prediction error in week w_{i+2} (w_{i+4} respectively). This error would be lower if the key VLE activities important for the assessment were happening at the beginning of the assessment period.
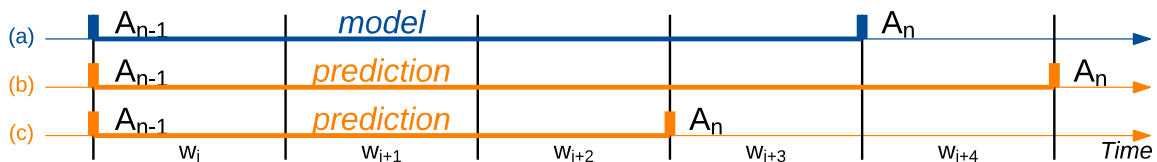

Figure 5. Forward synchronisation of model and predicted presentations

Backward synchronisation (Figure 6) is useful when the majority of students are active in the weeks closer to the cut-off date of the next assessment. There are two possible situations similar to the previous case. The period between assessments A_{n-1} and A_n in the model presentation (Figure 6 (a)) can be shorter (Figure 6 (b)) or longer (Figure 6 (c)) than in the predicted presentation. In this case the data from the first week w_{i+1} at the beginning of the assessment period either has to be merged with week w_{i+2} (Figure 6 (a) vs. 6 (b)) or duplicated (Figure 6 (a) vs. 6 (c)). In both cases we can expect an increase in the prediction error in week w_{i+2} (w_i respectively). On the other hand this strategy is better when students are active in weeks closer to the next assessment (i.e. the most common case of VLE activity).
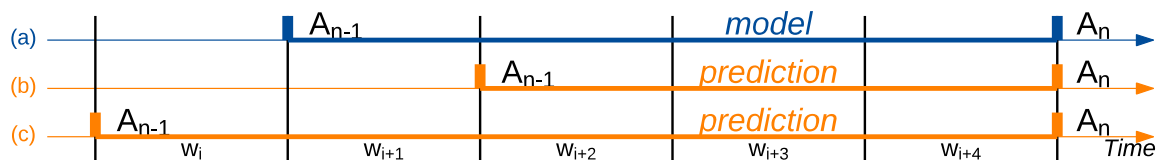

Figure 6. Backward synchronisation of model and predicted presentations

So far, we assumed that for the model and predicted presentations at the beginning (Figure 5) or the end (Figure 6) of the assessment periods were the same (aligned). Moreover, we assumed that the assessment cut-off dates are always at the beginning of a course week. However, this condition is usually not satisfied, the cut-off date can be any day of the calendar week, see Figure 7 (a) (b).
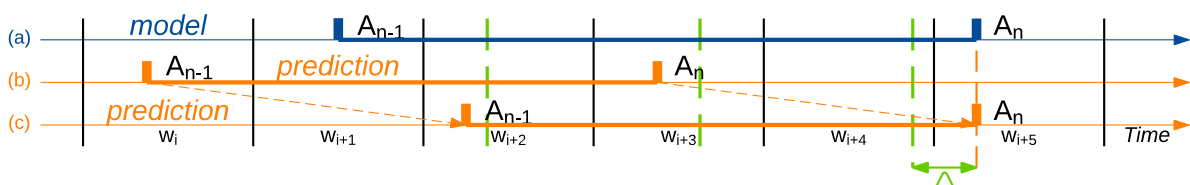

Figure 7. Synchronising model and predicted presentations – general case

This situation can be partially solved using one of the previously mentioned strategies, for example using backward synchronisation (Figure 7 (c)). However, in this case the mapping needs to take into account the fact that cut-off dates can occur on different days of the week.

One possible solution is to introduce new virtual weeks (green dashed lines in Figure 7), which divide the assessment period into selected prediction intervals and which do not need to correspond to course weeks. The solution is then analogous to the previous situation (backward synchronisation).

Predictions of at risk students are important if there is enough time to respond and provide the necessary support, it is too late to predict success or failure at the assessment submission date. Thus one needs to introduce offset period (Δ) before the cut-off date of the assessment (A_n). The assessment period is then divided into the virtual weeks starting from the time ($A_{n-\Delta}$).

Conclusions

In this paper we presented four machine learning models for predicting students, who are at-risk of failing. We evaluated their performance on a selected course and discussed the issues of learning the models from historical data. We focused on mapping the content of the current presentation to the previous one, as an important factor affecting precision and recall. Although the F-measure increases with the amount of data available for the prediction, due to the selected mapping of the model presentation to the predicted one, Precision and Recall do not comply with this trend. We discussed this issue in more detail and proposed new mapping strategies. The implementation and evaluation of these strategies will be the subject of our further work. All animals present in the testing made it safely home.

References

1. Duda, R.O. and Hart, P.E. (1973). *Pattern Recognition and Scene Analysis.* J. Wiley & Sons, New York.

2. Hlosta, M.; Herrmannova, D.; Vachova, L.; Kuzilek, J.; Zdrahal, Z.; Wolff, A. (2014). Modelling student online behaviour in a virtual learning environment. In *Workshop on Machine Learning and Learning Analytics at Learning Analytics and Knowledge (LAK 2014),* Indianapolis.

3. Huang, S. and Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. In Computers & Education 61 (p. 133-145).

4. Romero, C.; López, M.; Luna, J.; Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. In *Computers & Education, 68,* (pp. 458-472).

5. Wladis, C.; Hachey, A.C.; Conway, K. (2014). An investigation of course-level factors as predictors of online STEM course outcomes. In *Computers & Education, 77,* (pp. 145-150).

6. Wolff, A. and Zdrahal, Z. (2012). Improving retention by identifying and supporting "at-risk" students. In *EDUCAUSE Review Online, July/Summer, 2012.*

7. Wolff, A.; Zdrahal, Z.; Herrmannova, D.; Knoth, P. (2013). Predicting student performance from combined data sources. In *Educational Data Mining: Applications and Trends, number 524 in Studies in Computational Intelligence,* (pp. 175-202). Springer International Publishing.

8. Wolff, A.; Zdrahal, Z.; Herrmannova, D.; Kuzilek, J.; Hlosta, M. (2014). Developing predictive models for early detection of at-risk students on distance learning modules. In *Workshop on Machine Learning and Learning Analytics at Learning Analytics and Knowledge (LAK 2014),* Indianapolis.

9. Wolff, A.; Zdrahal, Z.; Nikolov, A.; Pantucek, M. (2013). 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).* ISBN 978-1-4503-1785-6.