ANALYSIS OF THE ACCURACY OF AN EARLY WARNING SYSTEM FOR LEARNERS AT-RISK: A CASE STUDY

David Baneres¹,², Abdulkadir Karadeniz²,³, Ana-Elena Guerrero-Roldán¹,², M. Elena Rodríguez-Gonzalez¹,², Montse Serra¹

¹Faculty of Computer Science, Multimedia and Telecommunications, Universitat Oberta de Catalunya (SPAIN)
²eLearn Center, Universitat Oberta de Catalunya (SPAIN)
³Open Education Faculty, Anadolu University (TURKEY)

Abstract

It is already well-known that technological developments have an undeniable effect on education. At most points, technological developments have a direct share in the improvement of the educational process. Early warning systems are one of these technological touches that affect education positively. Early warning systems, as an enhancement of learning analytics systems, are used to better support students based on their behavior and performance and identifying potential at-risk situations by collecting student data through technologies such as learning management systems or databases, which already have students previous signs of progress. Data, such as student participation, behavior and course performance constitute the basic input of an early warning system. Additionally, an early warning system does not require any special effort by teachers or any other participants rather than the existing data. It analyzes the risk status and achievement status of the participants for their future performance and presents them as a warning. This study aims to identify students at-risk by using the simple Gradual At-risk (GAR) predicting model in the Computer Structure course in the Universitat Oberta de Catalunya (UOC) and to provide early feedback based on the chance to pass the course. Computer Structure course with 249 enrolled students expands the knowledge of the hardware components in the undergraduate Bachelor of Computer Science. The course has four assessment activities (AA) during the semester timeline, and the early warning system is capable to identify potential at-risk students from the very first activity with an accuracy of the 73.49%. This study extends a previous one, which was aimed to develop an early feedback prediction system for learners based on data available in our institutional datamart (known as the UOC Datamart). The results of this study will demonstrate the effectiveness of the early warning system to identify at-risk students based on the GAR model and by using the Green-Amber-Red risk classification.

Keywords: Predictive models, at-risk student, early warning systems, online learning.

1 INTRODUCTION

Nowadays, Learning Management Systems (LMS) produce a vast amount of data during the learning process of learners. These data, which are related to the behavior of the students, conveniently processed, allows inferring valuable information to provide a personalized learning experience to students. Moreover, this information can be further analyzed and exploited to extract knowledge in the form of good practices and, therefore, create recommender systems for future students to complete their learning outcomes successfully.

Artificial intelligence is on the help of education. Predictive and classification models can be developed, and even rule-based recommendations can be automatically generated from these models for supporting learners and teachers. In this paper, we focus on a particular example of an intelligent system for education that is an early warning system (EWS). Those systems are mainly focused on supporting teachers in identifying at-risk students from the very beginning of the instructional process in the subject. Additional services can be added around the EWS to enhance its power. For instance, intervention measures such as automatic feedback messages or automatic recommendations.

Some examples of EWSs are early identification of at-risk students which may allow some type of intervention to increase retention and success rate [1][2][3][4], students dropout detection and reduction on face-to-face environments [5], students dropout on online settings [6][7][8], or students clustering to identify students with similar profiles [9][10].
In this paper, we analyze a case study related to an EWS developed at UOC. The analysis extends a previous experimental work presented on [11] where the performance of the EWS was analyzed on a first-year undergraduate subject. This work aims to observe whether the performance of the EWS is similar in another subject with different characteristics in terms of assessment activities.

The paper is organized as follows. Section 2 introduces the prediction models used within the EWS. Section 3 describes the EWS and its characteristics. Finally, the analysis of the EWS in the Computer Structure subject is presented.

2  PREDICTION MODEL

2.1 Gradual At-risk Model

In this section, we review the gradual at-risk model denoted as GAR model defined on [11]. First, let us define how the assessment process is performed at UOC. A continuous assessment model combined with summative assessment is applied to subjects. There are different assessment activities (AAs) during the semester and a final exam (FE) at the end of the subject. The final mark (FM) is computed based on a predefined formula for each subject where each assessment activity has a different weight depending on the significance of the assessment activity contents within the subject. It is relevant to know that the UOC grading system is based on qualitative scores on assessment activities. Assessment activity is graded with the following scale: A (very high), B (high), C+ (sufficient), C- (low), D (very low), where a grade of C- and D means failing the activity. In addition, another grade (N, non-submitted) is used when a student does not submit the activity. The grade for an activity depends on the evaluated mark according to the following table: A (9-10), B (7-8.9), C+ (5-6.9), C- (3.5-4.9) and D (0-3.4).

The GAR model is built for each subject, and it is composed of a set of predictive models defined as submodels based only on learner’s grades during the continuous assessment. A subject has a submodel for each assessment activity, and each submodel uses the grades of the current and previous graded assessment activities as features to produce the prediction. Note that, there is no submodel for the final exam since there is nonsense to produce a prediction when the final score of the subject can be computed straightforwardly from the complete set of grades (i.e., the assessment activities and the final exam).

The prediction outcome for the submodels is to fail the subject. This is a binary variable with two possible values: pass or fail. Note that, we are interested in predicting whether a student has chances to fail the subject, and we define this casuistic as an at-risk student. Although a global at-risk prediction is an interesting outcome, we focus on each subject to give simple messages to the student on which subjects she (or he) is at-risk.

Example 1. Let us describe the GAR model for a subject with four assessment activities. The GAR model contains four submodels:

\[
\begin{align*}
Pr_{AA1}(\text{Fail?}) &= \text{Grade}_{AA1} \\
Pr_{AA2}(\text{Fail?}) &= \text{Grade}_{AA1}, \text{Grade}_{AA2} \\
Pr_{AA3}(\text{Fail?}) &= \text{Grade}_{AA1}, \text{Grade}_{AA2}, \text{Grade}_{AA3} \\
Pr_{AA4}(\text{Fail?}) &= \text{Grade}_{AA1}, \text{Grade}_{AA2}, \text{Grade}_{AA3}, \text{Grade}_{AA4}
\end{align*}
\]

where \( Pr_{AA}(\text{Fail?}) \) denotes the name of the submodel to predict whether the student will fail the subject after the activity \( AA_{n} \) has been assessed. Each submodel \( Pr_{AA}(\text{Fail?}) \) uses all available grades (\( \text{Grade}_{AA1}, \text{Grade}_{AA2}, \ldots, \text{Grade}_{AA_{n}} \)), and it can be computed when the grade \( \text{Grade}_{AA_{n}} \) is available.

Note that, each submodel can be evaluated based on different accuracy indicators. We use three indicators:

\[
\begin{align*}
\text{ACC} &= \frac{TP+TN}{TP+FP+TN+FN} \\
\text{TNR} &= \frac{TN}{TN+FP} \\
\text{TPR} &= \frac{TP}{TP+FN}
\end{align*}
\]

where TP denotes the number of at-risk students correctly identified, TN the number of non-at-risk students correctly identified, FP the number of at-risk students not correctly identified, and FN the number of non-at-risk students not correctly identified. These indicators are used for evaluating the
accuracy of the complete model (ACC), the accuracy when identifying non-at-risk students (true negative rate - TNR), and the accuracy when detecting at-risk students (true positive rate - TPR).

The GAR model has two main advantages: 1) it is simple to apply and it can be adapted to any subject with a set of grades during the learning process, and, 2) it is simple to explain to students because it is only based on grading information.

2.2 Next Assessment At-risk Model

The GAR model only provides information about whether the student will fail the subject based on the last graded activity. To give more information to each student during her (or his) continuous assessment process, we predict the minimum grade that the student has to obtain in the next assessment activity to have chances to pass the subject by using the Next Assessment At-risk Model (NAAR) model.

This prediction is performed by running a submodel when all previous assessment activities have been graded. Then, all possible grades for the current assessment activity are explored to identify the minimum grade to pass the course.

Example 2. Let us take the submodel $Pr_{AA1}(\text{Fail}\,?)$ of Example 1. In order to know the minimum grade, six predictions are performed based on the possible grades the student can obtain in AA1. Each prediction will produce an output based on the chances to fail the subject. An output example is shown next based on the first assessment activity of the Computer Structure subject that will be analyzed in Section 4:

$$
\begin{align*}
Pr_{AA1}(\text{Fail}\,?) = (N) & \rightarrow \text{Output} = \text{Fail} \\
Pr_{AA1}(\text{Fail}\,?) = (D) & \rightarrow \text{Output} = \text{Fail} \\
Pr_{AA1}(\text{Fail}\,?) = (C-) & \rightarrow \text{Output} = \text{Fail} \\
Pr_{AA1}(\text{Fail}\,?) = (C+) & \rightarrow \text{Output} = \text{Fail} \\
Pr_{AA1}(\text{Fail}\,?) = (B) & \rightarrow \text{Output} = \text{Pass} \\
Pr_{AA1}(\text{Fail}\,?) = (A) & \rightarrow \text{Output} = \text{Pass}
\end{align*}
$$

where we can observe that learners tend to pass the subject when they get at least a B as a grade on the first assessment activity. Note that, this means that most of the learners pass when they obtain this grade. However, it is possible to pass the subject with a lower mark, but less frequently. Note that, this is the first assessment activity where there are no previous activities. On further activities, the grade of previous activities is also taken into account and the prediction is personalized for each student based on her grades.

3 EARLY-WARNING SYSTEM

In this section, we describe the design of the early-warning system briefly. An extended description with technical details can be found in [11]. The system is illustrated in Fig. 1 and it is composed of different services:
• Prediction service: The prediction service is responsible for generating predictions. Currently, NAAR model is used and it is trained independently for each subject to get more accurate results. The data are obtained from the institutional UOC datamart [12] where anonymized data from UOC campus are gathered. The information about grades from learners of the current semester are extracted from the Continuous Assessment Registry (CAR). The predictions (i.e., NAAR predictions) are stored on a temporal database ready to be consumed for other services.

• Dashboard service: This service provides the dashboards for instructors and learners providing different information based on the user role. In Section 3.1, we detail the dashboard for the learners.

• Feedback service: An additional service has been added to the system as an automatic intervention action based on the at-risk/not-at-risk situation of the student. We detail the service in Section 3.2.

3.1 Dashboard Service

The learner gets two types of information: information about the minimum grade to obtain in the next assessment activity and a global warning level indicator.

Based on the NAAR model, the learner gets the prediction in a stacked-like bar with only two outcome options: pass or fail. Figure 2 illustrates as an example the first assessment activity of the case study described in Section 4 and based on the predictions shown in Example 2. The x-axis shows the qualitative scores and the y-axis the name of the activity. Recall that the predictions were (Fail, Fail, Fail, Fail, Pass, Pass) for the respective scores (N, D, C-, C+, B, A), respectively. The stacked-like bar shows clearly the minimum grade the learner should obtain in the first assessment activity to have a high likelihood to pass the subject.

Using this stacked bar visualization, a dashboard has been designed to show the NAAR predictions for each learner. For each activity, the learner gets the information in two steps as it is shown in Fig. 3 and Fig. 4, respectively.
Before the submission of the assessment activity (see Fig. 3): At this point, the stacked bar informs about the minimum grade the student should obtain to get more chances to pass the subject. In the figure, we can observe how the stacked bar illustrated in Fig. 2 is integrated into the complete dashboard.

Figure 4. Learner dashboard after the submission of the first activity AA1.

After the assessment activity is graded (see Fig. 4): At this point, the stacked bar is updated with the grade that finally the learner obtained in the assessment activity (i.e., triangles above and below the mark). Then, a prediction for the next assessment activity is performed. Note that, this prediction is personalized for each student since it takes into account the grade already obtained for the learner in previous assessment activities (i.e., in this case only the AA1). In the figure, we can observe that the learner is assessed with a B score in the first assessment activity. Then, the dashboard generates a second prediction for the second activity (identified as AA2a). In this case, the EWS informs the learner that submitting the assessment activity has a positive impact in passing the subject whatever the grade is obtained (i.e., whether the learner puts some effort on performing the activity), although learners who did not submit the activity usually fails.

The dashboard also shows a warning level based on the Green-Amber-Red indicator. This type of indicator has been previously used in other early warning systems with a high appraisal [13]. In this case, the student is classified in three possible levels (i.e., green, amber, red) based on the chances to pass the subject. In our case, the green represents that the student is non-at-risk, while the red signal indicates a high likelihood to fail. The warning level is updated on every prediction showing to the student her (or his) current at-risk level. The signal is based on the decision tree illustrated in Fig. 5. As we can observe, the green signal is assigned when the prediction model has good accuracy in terms of TNR and the student gets a score equal or greater than the prediction. Similarly, the red signal is assigned when the student is not able to get the proper score, and the prediction model has a high accurate TPR. Note that, we consider that 70% of accuracy is the minimum value to consider an accurate model. Finally, the amber signal is assigned on non-accurate models.

As we can observe, there is a special black signal when there are more than two non-submitted activities. This color is only intended for teachers’ dashboards to identify potential dropout learners. Students continue seeing a red signal; meanwhile, teachers have another level of information granularity (dropout) and, thus, they can start additional intervention actions to try to reverse dropout.
3.2 Feedback Service

Based on Green-Amber-Red indicator, an automatic intervention action is performed. Learners get an automatic feedback email from the system. The contents of the feedback differ depending on the warning level. There are four different messages for each prediction (i.e. for each assessment activity). These messages are previously written by the instructor of the subject depending on the criticality of the assessment activity in terms of contents and the impact in the final mark. Meanwhile, students with a green signal get a comforting message with some small considerations to pass the subject, students with amber and red signals get a stronger message depending on the warning level. The message remarks the situation they are based on the chances to pass the subject and different recommendations for continuing the learning process such as additional exercises to perform, learning resources to access, or possible learning paths to pass the subject. Students with black color are treated distinctively since they are potential dropout learners. Additionally, to the recommendations of the red warning level, they are asked to contact the teacher for more help if there are interested. We observed that personal contact with a teacher in online learning could potentially revert the dropout.

4 EXPERIMENTAL RESULTS

The case study has been conducted in Computer Structure. This subject is in the specialty of Computer Architecture in the undergraduate Bachelor of Computer Science at the Universitat Oberta de Catalunya. Computer Structure expands the knowledge of the hardware components that have been studied on previous subjects. Also, this subject aims to practice the low-level language assembler by doing calls to assembler functions from C language. Note that, a high dropout rate of nearly 45% of enrolled students affects this subject.

The subject is composed of four assessment activities distributed during the semester where the AA2a and AA2b assessment activities are the first and second opportunity to submit the AA2. We selected this subject as the case study since it is composed of an assessment activity that can be submitted twice. We want to observe the behavior of GAR model on a subject with this type of assessment model.

The Final Mark (FM) is computed following the next formula:

$$FM = \text{MAX}(15\% \text{Grade}_{AA1} + 25\% \text{Grade}_{AA2} + 15\% \text{Grade}_{AA3} + 10\% \text{Grade}_{AA4} + 35\% \text{Grade}_{FE},$$

$$35\% \text{Grade}_{AA2} + 15\% \text{Grade}_{AA4} + 50\% \text{Grade}_{FE})$$

In this subject, the AA2 and the FE are the mandatory activities to pass. For this reason, there is a second opportunity to submit the AA2 in case of failing the first one. Related to the final exam (FE), it is placed at the end of the subject, and the EWS avoids producing a prediction since its minimum grade can be computed straightforwardly from the grade of the assessment activities and the minimum grade to pass the subject.
Table I. Performance GAR model in Computer Structure.

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>ACC</th>
<th>TNR</th>
<th>TPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>PrAA1</td>
<td>146</td>
<td>63</td>
<td>37</td>
<td>3</td>
<td>73.49</td>
<td>37.00</td>
<td>97.99</td>
</tr>
<tr>
<td>PrAA2a</td>
<td>113</td>
<td>5</td>
<td>95</td>
<td>36</td>
<td>83.24</td>
<td>95.00</td>
<td>75.84</td>
</tr>
<tr>
<td>PrAA3</td>
<td>147</td>
<td>32</td>
<td>68</td>
<td>2</td>
<td>90.20</td>
<td>68.00</td>
<td>98.66</td>
</tr>
<tr>
<td>PrAA2b</td>
<td>130</td>
<td>1</td>
<td>99</td>
<td>19</td>
<td>93.05</td>
<td>99.00</td>
<td>87.25</td>
</tr>
<tr>
<td>PrAA4</td>
<td>126</td>
<td>0</td>
<td>100</td>
<td>23</td>
<td>93.05</td>
<td>100.00</td>
<td>84.56</td>
</tr>
</tbody>
</table>

Table II. Performance of the Warning level indicator in Computer Structure.

<table>
<thead>
<tr>
<th>Warning Level *</th>
<th>AA1</th>
<th>AA2a</th>
<th>AA3</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>Fail</td>
<td>Pass</td>
<td>No.</td>
</tr>
<tr>
<td>B</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>R</td>
<td>109</td>
<td>76.58%</td>
<td>23.42%</td>
</tr>
<tr>
<td>Y</td>
<td>140</td>
<td>35.71%</td>
<td>64.29%</td>
</tr>
<tr>
<td>G</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Warning Level *</th>
<th>AA2b</th>
<th>AA4</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>Fail</td>
<td>Pass</td>
</tr>
<tr>
<td>B</td>
<td>105</td>
<td>98.13%</td>
</tr>
<tr>
<td>R</td>
<td>22</td>
<td>86.36%</td>
</tr>
<tr>
<td>Y</td>
<td>1</td>
<td>0.00%</td>
</tr>
<tr>
<td>G</td>
<td>121</td>
<td>9.09%</td>
</tr>
</tbody>
</table>

(*) The colours of the warning level are denoted by Black (B), Red (R), Yellow (Y) and Green (G).

Table I shows the accuracy of the GAR model in this subject. For each submodel, the true positive (TP), false positive (FP), true negative (TN), false negative (FN) and the accuracy indicators ACC, TNR, and TPR are summarized. We can observe how the accuracy indicators highly depend on the mandatory assessment activity AA2. The TPR is high in PrAA1 and PrAA3, while the TNR has a low value. These assessment activities are not mandatory, and we cannot affirm that a student will pass the subject based on the grade of these assessment activities, but dropout students are correctly identified as at-risk. A different casuistic is found in PrAA2a and PrAA2b, which correspond to the first and second opportunity to submit the AA2. Here, non-at-risk students are correctly distinguished by a TNR larger than 95%. Note that, the AA2 is mandatory and this is reflected in the predictive model.

Next, Table II summarizes the analysis of the warning level assignment. We summarize the number of students assigned to each level (No.) and the final performance of the students for each warning level (Fail, Pass) giving insights about the correct assignment. The accuracy of the TNR and TPR of the submodels have a big impact on their assignment. The Green color cannot be assigned on the AA1 and the AA3 because the TNR is smaller than the threshold of 70%. The Black color correctly identifies dropout students at all levels, but the red one has some difficulties to identify at-risk students in the AA2a and AA3 assessment activities correctly. Some students pass the subject due to the second opportunity, and this is shown in AA2b.
The statistical significance of the improvement on the final score distribution has also been checked. We used the unpaired two-sample Wilcoxon test due to the non-normal distribution of the final scores. Here, we assume as the null hypothesis that the scores are worse or equal in the previous semester. The p-value < 0.02 and the null hypothesis can be rejected. Note that, the median of the final mark increases from 7.1 to 7.2 and the dropout rate decreases from 43% to 40%. The results regarding retention and score are slightly better, but they cannot be only inferred from the utilization of the EWS.

Finally, the opinion of the students is shown in Table III. When the dashboard is updated with the assessment activity grade (see Fig. 4), the student has the chance of answering a survey composed of three questions: 1) do you think that this prediction is useful?, 2) are you going to continue the subject?, and 3) what is your mood after the prediction?. A Likert scale from 1 to 5 was used being 1 a strongly negative answer and 5 a strongly positive one. Note that, the questions are answered within the student’s dashboard and, thus, the questionnaire is not anonymous. This helps to correlate opinions with the assigned warning level. We received from 150 answers in the AA1 to 73 in the last assessment activity (AA4). The error margin at the end of the subject reaches 10% in a 95% interval confidence. Although we cannot generalize these results to the whole subject, interesting insights can be observed.

The question related to prediction usefulness has a good appraisal on all levels except the Red one. A value larger than 75% is observed in the rest of the levels. The continue appraisal decreases depending on the warning level having a smaller value on students at the Black color. It is interesting to observe that its value is still high in the AA3. Teachers assume that students that will apply in the second opportunity of the AA2 are motivated to continue. Unfortunately, the value of all indicators of at-risk students in the AA2b and the AA4 decreases since these students failed the mandatory assessment activity AA2 and, thus, they will probably fail the subject.

5 CONCLUSIONS

In this paper, we have presented a case study based on an EWS developed at Universitat Oberta de Catalunya. This case study extends the case study presented in [11]. Similar results were obtained in terms of accuracy of the GAR model, performance of the warning level classification and appraisal of the learners.

The presented GAR model has a high accuracy rate in the analyzed subject. One disadvantage of the model is that it is not continuously updated since it depends on the continuous assessment activities. As future work, we are planning to improve this model in terms of accuracy and update time. The first improvement is focused on having more accurate models by adding other relevant data of learners. The second improvement is focused on providing better personalized follow-up feedback to learners during
their learning process between the assessment activities submission and to limited to the end of the activity.

ACKNOWLEDGMENTS

This work was funded by the eLearn Center at Universitat Oberta de Catalunya through the project: New Goals 2018NG001 “LIS: Learning Intelligent System”.

REFERENCES


