Predicting Student Success in a Blended Learning Environment

Steven Van Goidsenhoven
KU Leuven
Belgium
steven.vangoidsenhoven@kuleuven.be

Daria Bogdanova
KU Leuven
Belgium
daria.bogdanova@kuleuven.be

Galina Deeva
KU Leuven
Belgium
galina.deeva@kuleuven.be

Seppe vanden Broucke
KU Leuven
Belgium
seppe.vandenbroucke@kuleuven.be

Jochen De Weerdt
KU Leuven
Belgium
jochen.deweerdt@kuleuven.be

Monique Snoeck
KU Leuven
Belgium
monique.snoeck@kuleuven.be

ABSTRACT

Blended learning is gaining ground in contemporary education. However, studies on predictive learning analytics in the context of blended learning remain relatively scarce compared to Massive Open Online Courses (MOOCs), where such applications have gained a strong foothold. Data sets obtained from blended learning environments suffer from a high dimensionality and typically expose a limited number of instances, which makes predictive analysis a challenging task. In this work, we explore the log data of a master-level blended course to predict the students' grades based entirely on the data obtained from an online module (a small private online course), using and comparing logistic regression and random forest-based predictive models. The results of the analysis show that, despite the limited data, success vs. fail predictions can be made as early as in the middle of the course. This could be used in the future for timely interventions, both for failure prevention as well as for reinforcing positive learning behaviours of students.

CCS CONCEPTS

• **Applied computing** → **Distance learning**; *E-learning*; • **Mathematics of computing** → **Regression analysis**; *Exploratory data analysis*; • **Information systems** → *Content analysis and feature selection.*

KEYWORDS

blended learning, grade prediction, logistic regression, random forest classification, machine learning, e-learning, learning analytics, feature extraction

ACM Reference Format:
Steven Van Goidsenhoven, Daria Bogdanova, Galina Deeva, Seppe vanden Broucke, Jochen De Weerdt, and Monique Snoeck. 2020. Predicting Student Success in a Blended Learning Environment. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK '20), March 23–27, 2020, Frankfurt, Germany*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3375462.3375494

1 INTRODUCTION

Blended learning environments, which "involve a mix between online and face-to-face teaching" [Oliver and Trigwell 2005, p. 3], provide a useful source of information on student behaviour, even though the log data produced by blended courses' online modules may not be as rich as the data obtained from Massive Open Online Courses (MOOCs). In this work, we analyse the log data generated in a master-level blended course, and 1) evaluate the possibility of predicting the success of a student in a blended learning environment, 2) explore the links between the type and timing of students' online activities and the outcome of the course, and 3) identify, from which period in time a student's outcome can be predicted accurately.

The main contribution of this paper is the demonstration that, based on the logged data of the behaviour of students within the online module, one can built accurate classification models for predicting student success. Moreover, we show that clever data aggregation is required in order to supply predictive analytics models with high-quality features to base their predictions on, especially when confronted with smaller data sets. Furthermore, the predictive models learnt provide interesting insights regarding the possible reasons why students pass or fail. Finally, this study also demonstrates that predictive models can be successful early on in the course, which gives further evidence that intervention strategies based on these predictive models are worthwhile. The results of the analysis show that for the particular course, success vs. fail predictions could be made as early as in the middle of the course, which could be used in the future for timely interventions, both for failure prevention and for supporting the positive learning behaviour of students.

The paper is structured as follows: Section 2 presents the state of the art in blended learning data analytics and failure or dropout prediction methods, in Section 3 the context, data set and methodology are described, and the results of the analysis are shown in Section 4. Section 5 discusses the results, as well as the limitations of the study, and Section 6 concludes the work and provides several possible directions for future research.

LAK '20, March 23–27, 2020, Frankfurt, Germany

ACM ISBN 978-1-4503-7712-6/20/03…$15.00
https://doi.org/10.1145/3375462.3375494

2 STATE OF THE ART

In this section, we discuss related work focusing on data analytics in blended learning contexts and student success prediction.

2.1 Blended learning data analytics

The recent two decades showed a rise of various forms of "flexible" education, ranging from purely distant, online learning to blended learning or employment of flipped classrooms. Given the rise in interest for such forms of learning, more and more educational institutions switch from traditional to blended learning environments and employ Virtual Learning Environments (VLE), such as edX, Coursera, Codecademy and others, in their everyday educational practice. These new practices offer "advanced collaboration and communication, convenience (costs, didactics, learning) efficiency, VLE user control, personalisation, ubiquity, task orientation and timeliness of VLE driven learning and teaching." [Mueller and Strohmeier 2011, p. 1] Despite the fact that these virtual platforms keep detailed information about the study behaviour and activity of the students, and that MOOC data analysis has gained particular interest in the field of learning analytics, blended learning environments, as well as Small Private Online Courses (SPOCs) remain relatively underrepresented in studies related to learning analytics. One of the possible reasons that can explain this research gap is that blended learning generates much smaller data sets due to the small number of students (compared to MOOCs) and the fact that a significant part of the educational process, i.e. the face-to-face aspect, is not logged.

From an educator's perspective, blended learning has a significant value compared to both traditional and purely online learning. Access to the logs of the students' studying behaviour from the online platform offers the opportunity to use learning analytics (LA) techniques to find patterns in the studying behaviour that could give valuable insights into the student's potential achievements, as well as the actual effectiveness of the course. Online modules of blended courses usually produce the types of data similar to the ones produced by MOOCs, only for a smaller number of students. The main types of data are: logs of student activity (including various actions performed by a student), online assessment grades and social interaction data, such as comments and posts on a platform's dedicated forum. At the same time, those aspects that were not caught by the log data could be discovered during face-to-face sessions, thus, more refined adjustments can be made in the educational process. According to [Zacharis 2015, p. 1], "the analysis and interpretation of tracking data of students' activities online should be a seamless part of a blended learning classroom workflow". Nevertheless, in reality the "seamlessness" is hardly ever achieved, and very few data-driven recommendations to the teachers of blended learning courses are available, especially regarding the dropout, academic failure prediction and possible points for interventions. Institutions of higher education, in general, rarely realise the full potential of the learning data stored by its student information systems and learning management systems [Dawson et al. 2010]. Even though the student data missing from the online platform logs could be enriched by the multi-modal data (e.g. pulse or context of activity) collected during students' self-regulated activities [Di Mitri et al. 2017], the complexity of the multi-modal learning analytics solutions hinders the wide use of those by educators [Rodríguez-Triana et al. 2018]. Thus, the potential advantage of a blended learning setting laying in the "offline" access to the students often remains unrealised, and the data collected by the VLEs becomes the main, and often only source of student data.

2.2 Predicting academic achievement in a blended learning setting

The problem of student failure or dropout remains one of the main challenges contemporary educators face, especially in the higher education setting. Works on detection of students at-risk range from large campus-wide studies taking into account various types of general success factors, including demographics or prior education [Hoffait and Schyns 2017] to predictive studies for a particular blended or online course [Costa et al. 2017]. The prediction of students' grades in blended learning environments is not an easy task. The nature of the blended learning environment makes it difficult to collect concise and full data sets, as not all students are in need of the full capabilities of the online course to successfully pass the exams: students often have higher preference for face-to-face over online learning [Paechter and Maier 2010], and if the online part of the course is not compulsory, some of them may choose to study the materials in a different setting. Since machine learning algorithms are quite challenging to implement on data sets with incomplete data, this paper tries to show if a certain way of modelling and pre-processing of the data can help to mitigate these data issues related to Blended Learning Environments (BLE). Another big issue that is related to these BLEs is the small sample size of the data set, because of the tendency to be only given to a selected number of students. In addition to their low amount of data points, BLEs tend to collect high amounts of data features, comparable to MOOCs. For this reason, the data sets of BLEs show similarities with high dimensional, low sample size data sets. These types of data sets tend to be susceptible to the "curse of high dimensionality"[Sarkar and Ghosh 2019; Weber et al. 1998]. These concerns may position the task of predicting attainment with data analytics instruments as not feasible for BLEs: according to [Picciano 2014, p. 42], "at the present time <...> these software are best suited for fully online environments, not face-to-face or blended learning environments".

Despite all the difficulties mentioned above, a number of works undertakes the challenge of predicting academic success based exclusively on BLE data, using various combinations of learning data and LA approaches. Using the combination of student interaction, attendance and grade, the authors of [Harrak et al. 2018] proposed a profiling approach based on the types and quantity of questions asked by students on an online platform, employing K-means clustering to map the students according to the typology of their questions. Very few papers are found to report successful predictive models based exclusively on student actions log data within a BLE, without taking the contents of social interaction or self-reported student data into account. In [Zacharis 2015], an analysis of students' tracking data from a learning management system (LMS) is conducted to find the significant correlations between certain types of student activities and course grade and build a predictive model to identify the students at risk. After a step-wise multiple regression analysis, the authors found that the final student grade could be predicted by four variables: reading/posting messages, content creation contribution, quiz efforts and files viewed. After a classification using binary logistic regression, the model reached the accuracy of 0.813 (10-fold cross-validation). A larger-scale study by [Nguyen et al. 2018] also explored the student activity data from an LMS forecasting the student interactions and achievement of learning outcomes at different stages of the course (by time periods of several weeks) using linear regression. This was done to provide students with timely information about their performance in the form of a dashboard, as a measure for early intervention to prevent failure. The presented model predicted student failure with an accuracy of 0.70. These few examples show that investigating the data from the online part of a blended course has some potential, but at the same time that there is a lot of room for further exploration.

3 DATA SET AND METHODOLOGY

The data set used for this study comes from the mandatory course "Architecture and Modelling of Management Information Systems" taught, amongst others, within the Master programme Information Management at the Faculty of Economics and Business at KU Leuven. This course ran in the second semester of the 2017-2018 academic year and followed a blended learning approach. The course was composed of the following elements: live lectures, eight exercise sessions, a group assignment, a supporting online course (SPOC structure) and a final written exam. The online course was created on the edX Edge platform with web lectures and online quizzes. Extra supporting tools such as forums were also available in the online platform. However, as the course was partially given on campus, most of the students preferred the private social network groups. Therefore the data from these supporting tools were not accessible for the purposes of this study.

The data extracted from the edX Edge platform was captured from the start of the course, mid February 2018, until the exam on the 15th of June 2018. The raw data was exported from the edX Edge platform in a nested JSON (JavaScript Object Notation) format, which contained various information identifying student activities within the platform, such as event type, its timestamp and a reference link to the corresponding course page, and the information related to students' identity, such as username, user device and IP address. After the raw data was preprocessed and unnested, any information that can possibly reveal student identity was removed, and the username was anonymised. As a result, the data set after the first round of preprocessing contained the values as summarized in Table 1: Time, Session, Referrer, Event type and Anonusername.

The data was then further processed with the Python Pandas library to a more workable format. Chapters and sections were extracted from the Referrer field and the events were formatted into a more readable format. Furthermore the data set was merged with the information containing the grades of the students. To identify students who failed or passed their final exam, an extra label 'passed' was added. Students who had a score lower than 10/20 got a value of 0 and students who had a grade of at least 10/20 received a value of 1.

3.1 Outliers

In order to ensure that the results would not be impacted by potential outliers and anomalies, two different detection methods were applied. First we statistically analysed the data by using a box plot of the total event counts for all the students which resulted in the detection of two potential outliers. Moreover, the same outliers were confirmed through the use of an isolation forest [Liu et al. 2008], using the scikit-learn library [Pedregosa et al. 2011] in Python. Finally, these two outliers were both removed from the data set before any further analysis was done.

3.2 Predictive modelling data preparation

In an effort to better prepare the data for use in predictive modelling, further processing of the data was done. More specifically, for every event type (as listed in Table 2), activity counts for each of the students for each of the 18 weeks were produced. Hereto, we counted activity level per day (number of occurrences of a particular event type), and then averaged this for every week. Subsequently, these weekly averages were again averaged over different time periods. We decided to produce three different data sets, by separating the weeks into respectively 5, 9 (biweekly) and 18 (weekly) periods. A detailed example of the aggregation into 5 periods can be found in Table 3. Standardization was applied to all features.

3.3 Predictive models

To decide which predictive modelling technique would be most appropriate, a few concerns were taken into account. In the first place, models for student success prediction need to have good explainability, given that both teachers as well as experts would like to understand why a particular prediction was made, especially also when considering intervention strategies. In addition, we would prefer the techniques to be quite robust against overfitting, due to the fairly high number of features compared to the number of observations. Please observe that exactly because of the low number of observations in the data set, we transformed the problem into a binary classification instead of predicting an exact grade (which would boil down to a regression). All this considered, in this paper we opt for two predictive modelling techniques: logistic regression and random forests. The former aligns very well with our concerns listed above (explainability and robustness against overfitting), while the latter modelling technique is considered one of the most powerful classification techniques available nowadays, and thus serves as a solid benchmark for the logistic regression. In addition, it is well-known for its ability to define feature importance values [Saeys et al. 2008], which could be used in helping lowering the high dimensionality in terms of the number of features. The two predictive modelling techniques were applied to the data with the scikit-learn module in Python[Pedregosa et al. 2011]. The hyperparameters of the models were defined by using a grid search algorithm, using the leave-one-out cross-validation (LOOCV) accuracy as scoring parameter and can be found in Tables 4 and 5. Finally, a visual confirmation of the prediction model was built using PCA in an attempt to visually validate the results from the prediction models.

Table 1: Elements of the data set obtained after the first round of preprocessing of the native edX JSON format

Value	Explanation
Time	The timestamp of the recorded event.
Session	The session ID generated by the edX platform.
Referrer	The hyperlink at which page the event took place
Event type	The name of the event
Anonusername	The hashed username of the student for anonymity reasons.

Table 2: Overview of all event types in the data set

Event Type	Explanation
ADD A POST	User creates new thread
ADD USER	User is added
CLICK BOOKMARK	User used bookmark previously manually added
CLICK COURSE TOOL HEADING	User clicks on link in the course tool heading
CLICK NAVIGATION BAR	User selects any tab in the unit navigation bar to navigate through the course
CLICK RESUME	User resumes course where last left off
LOAD VIDEO	When video is fully rendered and ready to play
NEXT PAGE	User clicks in navigation control to go to the next page
PAGE CLOSE	User closes the page
PAUSE VIDEO	User pauses the video
PLAY VIDEO	User starts to play the video
PREVIOUS PAGE	User clicks in navigation control to go to the previous page
RESET PROBLEM	User resets the answer to a problem
SAVE ANSWER OF A PROBLEM	User saves the answer to a problem
SEEK VIDEO	User seeks within a video
SELECT LINK	User clicks any hyperlink within the course content
SERVER GRADES SUBMITTED	Server logs event when user submits and successfully saves an answer
SERVER PROBLEM CHECKED	Server logs event when user checks problem
SHOW ANSWER OF A PROBLEM	User clicks "show answer" of a problem
SIDEBAR UPSELL DISPLAYED	NA
SPEED CHANGE VIDEO	User changes playing speed of video
SUBMIT PROBLEM	User submits a problem
USER ENROLLEMENT	Logged when user is successfully enrolled
VIDEO COMPLETED	Logged when user successfully finishes a video
VIEW FORUM	User views a thread on the forum

Table 3: Detailed overview on how the 5 periods are defined in the paper

Period	Weeks included	Course in practice
1	1-4	The start of the semester, first homework.
2	5-8	Second homework, beginning of Easter holidays.
3	9-12	End of Easter holidays, the group work task preparation.
4	13-16	Finalizing the group work task, course wrap-up.
5	17-18	Preparation period before exams.

3.4 Classifier performance evaluation

To evaluate the classification models, some performance metrics were calculated. Considering the small amount of data points in our data set, this paper used the leave-one-out cross-validation accuracy as the base evaluation method to compare the two different modelling techniques [Airola et al. 2009]. Leave-one-out cross validation calculates the mean accuracy over 61 different runs, in which it uses 60 data points to predict the left out data point as the test value. Every run uses a different "left-one-out" data point, until all data points have been used once as a test value. In addition to the LOOCV accuracy, we opted to also include the AUC score of the model, the confusion matrix, also known as the error matrix [Stehman 1997], the precision, recall, f1-score, accuracy, macro average and weighted average [Herlocker et al. 2004; Witten et al. 2011].

Table 4: Hyperparameters from the logistic regression model

Hyperparameter	Value
Norm used in the penalization	l2
Inverse of regularization strength (C)	0.21
Algorithm used in the optimization problem	SAGA
Fit intercept	True

Table 5: Hyperparameters from the random forest classification model

Hyperparameter	Value
Function to measure a split	Gini impurity
Maximum features to select	10
Number of estimators	50
Maximum depth	9

3.5 Lowering the high dimensionality

For the reason that the "curse of high dimensionality" could also influence our predictive models, the first priority in further fine tuning the logistic regression model was lowering the high dimensionality (amount of features), compared to the low sample size of the data set. With all the different events included, the base data set consisted out of 25 different event types (see Table 2) and 18 weeks, resulting in a total of 450 features. Even when the data set would be aggregated into 5 periods, instead of all the 18 weeks, it would still have 125 features. To retain only the most influential events, we applied a manual backward feature selection by identifying those features, using both models, which had very little to no influence on the model's predictions. In addition to the feature importance results from the models, contextual knowledge about the course was used to exclude some events. Given both processes, the final remaining events, in descending order of importance, were: "SUBMIT PROBLEM", "SHOW ANSWER TO A PROBLEM", "CLICK NAVIGATION BAR", and "PLAY VIDEO". Altogether this resulted in three different data sets with a final dimensionality of 61 data points and either 20 (for 5 periods), 36 (for 9 periods), or 72 features (for 18 periods), originating from the 4 retained events.

4 RESULTS

4.1 Visual descriptive statistics

Figure 1 shows the distribution of the grades and the passed label of the students of the course. The passed students outnumber the failed students with a relative frequency of 0.64 for the passed students and 0.36 for the failed students. Remark that the high amount of grades between 0 and 2 stems from the fact that 7 students didn't show up on their exam and consequently got a grade of 0. Figure 2 shows the general distribution of the grades of the students, coloured by the passed label. This figure further emphasises the skewness of the passed label distribution in general and the skewness of the grades within the failed students group.

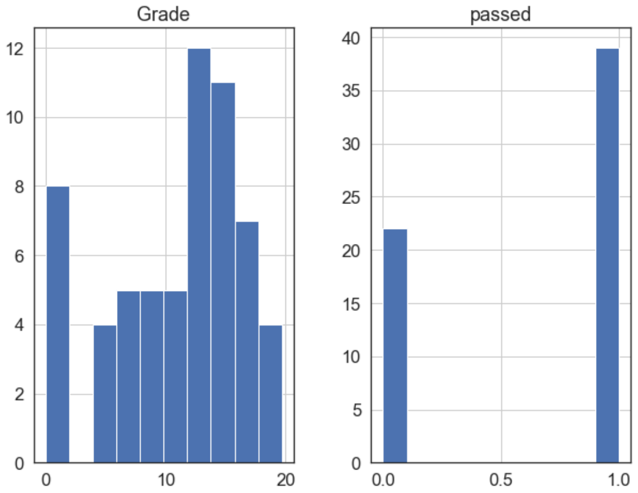

Figure 1: Histogram of the students' grade and passed label

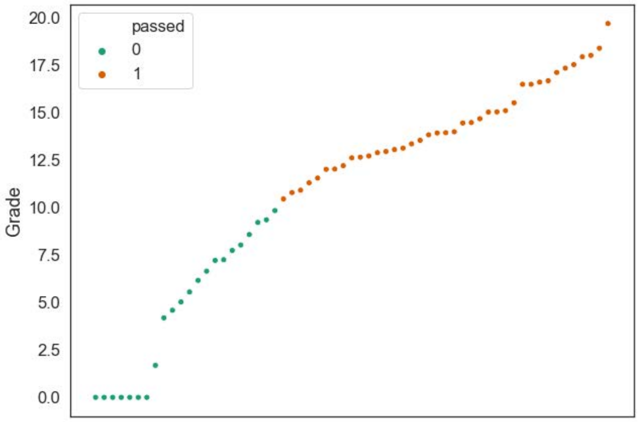

Figure 2: Visual overview of students' grades in a scatter plot

Table 6: Comparison of the predictive performance of the logistic regression models according to the number of periods considered

Nr. of periods	AUC	LOOCV accuracy
5	0.88	0.84
9 (bi-weekly)	0.92	0.84
18 (weekly)	0.96	0.77

4.2 Predictive modelling

When comparing the differently aggregated data sets, as seen in Table 6, we can notice a possible trend of overfitting. This could be explained due to the rising AUC score, while the leave-one-out cross-validation score is the same or even worse. To mitigate potential overfitting, and to favour a lower dimensionality of the data set to avoid the potential influence of high dimensions on the models, all calculations have been done with the data set aggregated into 5 periods, except when otherwise stated.

Table 7: Additional performance metrics per class label for the Logistic regression model trained on the full 5-periods data set

label	precision	recall	f1-score	support
0	0.90	0.82	0.86	22
1	0.90	0.95	0.92	39

When both models were tested on the 5-periods data set, the leave-one-out cross-validation accuracy of the logistic regression model amounted to the constant value of 0.84. On the other hand, the leave-one-out cross-validation accuracy of the random forest classifier amounted to a inconsistent value between 0.75 and 0.80 after multiple runs. It is interesting that both models show fairly high LOOCV accuracy values. Due to the fact that the logistic regression model has a higher LOOCV accuracy value, seems to perform more stable and has the ability to give more details about the predictive process, we believe that this model, compared to the random forest classifier, is the most interesting model to use within this paper. Consequently all further analysis has been focused on the logistic regression model.

The metrics of the logistic regression model when trained and validated on the full 5-periods data set can be found in Table 7. When looking at the metrics of the logistic regression model, a few particularities can be distinguished. First, there is a difference between the recall score of the failed students and the passed students. This same difference can be observed in the confusion matrix in Table 8. This suggests that correctly predicting failed students in this data set tends to be more difficult than predicting the passed students.

Figure 3 shows a line plot of the leave-one-out cross-validation accuracy for the logistic regression model, ran on the 5-periods data set, with the cumulative periods. Here it is clear that in the first period the predictions are not optimal because the accuracy is very close to the ratio between failed and successful students. Information from the second period added to the first period makes the model more accurate. When information from the third period is added, the accuracy reaches its highest point of 0.85. Adding more periods to the third period even slightly lowers the leave-one-out cross-validation accuracy.

Table 9 and 10 show the full results from the logistic regression prediction model, ran over the 5-periods data set for each of the students. Table 9 contains all the predictions for the students who failed on their exam, while Table 10 contains all the predictions for the students who passed their exam. The pred-column represents the binary prediction results from the logistic regression model. The passed column represents the true values of the students exam result which they passed or failed. The result column shows if the model made a wrong (0) or correct (1) prediction. Finally, the two prob_-columns show the probability of the prediction on a scale from 0 to 1. Both tables show the results from the wrongly classified predictions in the first rows, which are also shown in the confusion matrix in Table 8. Remarkably, these values show that 2 out of the 4 False Positives are very indecisive.

A detailed overview of the importance of the different features, defined by the logistic regression model ran over the 5-periods data set, can be found in Table 11. The importance of every feature is sorted by the absolute value of the coefficient in descending order. Every row in the table is a different feature, split up into two columns: the "period" column represents the period in which the events were aggregated and the "event type" column represents the type of event by which it was aggregated. This table shows that period 2 was a very influential period during the course, but only regarding submitting problems and viewing the answers to said problems. Interestingly, playing videos in that same period, showed little influence in the model. Likewise, submitting problems in the first period of the course seems to be hardly predictive. Furthermore it is shown in Table 11, that all events, except "CLICK NAVIGATION BAR", are very widely divided between all 5 periods. Therefore defining an influential period or event type, without the other, is a very difficult task.

Table 8: Logistic regression confusion matrix

	Predicted 0	Predicted 1
Actual 0	18	4
Actual 1	2	37

Finally, Figure 4 shows a PCA plotting of the 5-periods data set. The coloring is based on the actual data from the passed label in the data set. The green colour identifies all the students who failed their exam and the orange colour all the students who passed their exam. All the crosses identify the correct predictions from the logistic regression model and the circles identify the wrongly predicted students. In this figure the clusters of the successful and unsuccessful students can be clearly identified, with a transition zone in the middle. The two students who were wrongly predicted as bad students, as can be seen in Table 8 and Table 9, are both in the area of the bad students. Three of the 4 wrongly predicted good students (see Table 8 and Table 10), but who actually failed on their exam, are clearly in the area of the good students. Only one student who was wrongly predicted by the model is in the transition zone.

5 DISCUSSION AND LIMITATIONS

Despite of the limited size of the data set, the context of the blended learning environment and its high dimensionality, we managed to create an accurate predictive model to see if a student will fail or pass on their final exam. More precisely, it managed to predict the success rate of the students with a leave-one-out cross-validated accuracy of 0.84, which is in line with the LOOCV accuracy of the random forest classifier. We believe that these results were obtained by carefully lowering the high dimensionality of the data set, without removing important data for the predictability of the models.

The model managed to reach an *acceptable* accuracy level (LOOCV accuracy of 0.75) starting from data up to the second out of five total periods. This is equal to data up to the 8th week, out of 18 weeks of the course. This period of 8 weeks consists of 7 out of the 13 lecturing weeks and the first week of the Easter break of 2 weeks. In addition, the model achieves *high* accuracy (LOOCV accuracy of 0.84) starting from data up to the third out of five total periods. This is equal to data up to the 12th week, out of 18 weeks of the course. This period of 12 weeks consists of 10 out of the 13 lecturing

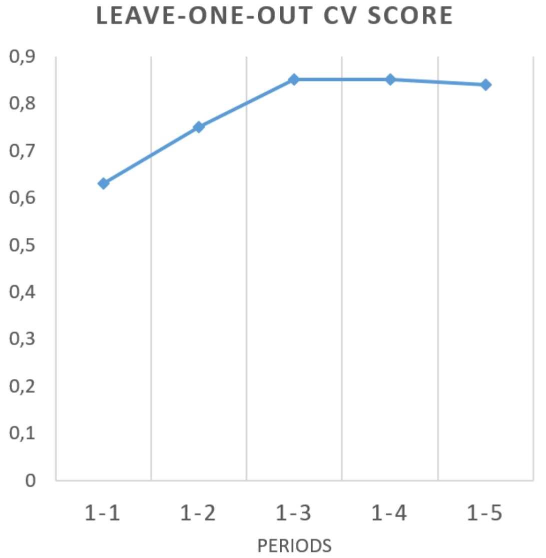

Figure 3: Line plot of the LOOCV accuracy values of the logistic regression models obtained for the different cumulative periods

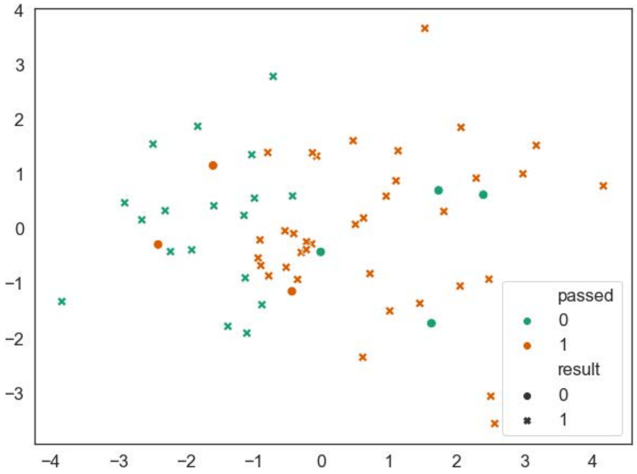

Figure 4: PCA plotting of the 5-periods data set into 2 dimensions

weeks and the Easter break of 2 weeks. As a result, this creates the possibility to warn students well before the completion of the course. Besides the potential to identify possible failing students, it also has the quality to identify students who have a high chance of passing the exam, as predicted by the model, but did not on their exam. In this particular case, three out of the four misclassified students (i.e. the first four student from Table 9) passed the course in the resit exam period. Provided that other sources of data regarding the students of the blended learning course can be accessed, further research could be done regarding the misclassified students. Models like these could at least help in identifying students with the potential to pass the course, but who are in need of extra assistance (e.g. like the first student in Table 9 who did not attempt the exam in the first exam session).

Table 9: Full prediction results from the logistic regression model for all students who failed the exam

id	pred	grade	passed	result	prob_0	prob_1
55	1	0.00	0	0	0.11	0.89
58	1	9.35	0	0	0.22	0.78
22	1	7.21	0	0	0.44	0.56
20	1	8.58	0	0	0.50	0.50
8	0	6.65	0	1	0.51	0.49
46	0	7.25	0	1	0.51	0.49
30	0	6.16	0	1	0.55	0.45
3	0	0.00	0	1	0.56	0.44
54	0	5.03	0	1	0.61	0.39
44	0	8.03	0	1	0.63	0.37
25	0	7.74	0	1	0.63	0.37
35	0	5.55	0	1	0.68	0.32
18	0	0.00	0	1	0.70	0.30
32	0	4.19	0	1	0.70	0.30
47	0	0.00	0	1	0.72	0.28
23	0	9.21	0	1	0.72	0.28
1	0	0.00	0	1	0.73	0.27
10	0	9.84	0	1	0.76	0.24
24	0	4.59	0	1	0.76	0.24
52	0	0.00	0	1	0.77	0.23
48	0	0.00	0	1	0.78	0.22
59	0	1.69	0	1	0.81	0.19

Despite the small difference between the leave-one-out cross-validated score of the logistic regression model and the random forest classifier, we believe that the logistic regression model has the best potential in predicting the success rate of students in a blended environment. Especially when, like in this case, a very small and limited data set is used. Moreover, the logistic regression by default provides a reliable prediction probability, which is a useful piece of information to obtain in this context.

Finally, the relative importance of the different variables provides useful feedback to the teacher as well, especially when looking at the connection between the event type and the period in which it was recorded. Table 11 shows that it is very difficult to define the importance of periods or event types without considering their inter-dependence. In this specific case the problem in the second period of the course corresponds to what is known as a hard part of the course that needs to be mastered in order to master the remainder of the course, whereas the problems offered in the first part of the course seem to have no predictive value. In this sense the "learning analytics" is at the same time also "teaching analytics" as it is able to confirm or challenge the effectiveness of a course's design.

When comparing our study with other studies related to BLEs, as for example done by [Harrak et al. 2018], it becomes apparent that our study did not have access to data related to students asking questions on an online platform, their attendance, their interaction, or other BLE related data. It is therefor quite difficult to compare our

Table 10: Full prediction results from the logistic regression model for all students who passed the exam

id	pred	grade	passed	result	prob_0	prob_1
41	0	13.35	1	0	0.72	0.28
7	0	10.45	1	0	0.63	0.37
14	1	12.03	1	1	0.49	0.51
28	1	16.68	1	1	0.43	0.57
34	1	17.34	1	1	0.38	0.62
6	1	17.11	1	1	0.38	0.62
39	1	10.91	1	1	0.33	0.67
33	1	13.13	1	1	0.32	0.68
51	1	14.45	1	1	0.30	0.70
2	1	12.95	1	1	0.26	0.74
50	1	12.71	1	1	0.25	0.75
17	1	12.89	1	1	0.24	0.76
21	1	13.94	1	1	0.22	0.78
9	1	12.61	1	1	0.22	0.78
16	1	19.69	1	1	0.20	0.80
57	1	12.01	1	1	0.20	0.80
40	1	12.65	1	1	0.19	0.81
31	1	15.10	1	1	0.19	0.81
60	1	11.30	1	1	0.17	0.83
38	1	16.49	1	1	0.17	0.83
43	1	13.54	1	1	0.17	0.83
29	1	10.79	1	1	0.17	0.83
37	1	13.83	1	1	0.16	0.84
53	1	17.53	1	1	0.16	0.84
42	1	18.39	1	1	0.16	0.84
4	1	18.01	1	1	0.15	0.85
5	1	13.93	1	1	0.15	0.85
27	1	15.51	1	1	0.15	0.85
45	1	12.20	1	1	0.14	0.86
13	1	11.55	1	1	0.14	0.86
56	1	13.05	1	1	0.12	0.88
36	1	17.94	1	1	0.12	0.88
15	1	14.48	1	1	0.11	0.89
12	1	14.68	1	1	0.11	0.89
11	1	16.49	1	1	0.09	0.91
49	1	16.60	1	1	0.09	0.91
26	1	13.99	1	1	0.07	0.93
19	1	15.03	1	1	0.04	0.96
0	1	15.04	1	1	0.02	0.98

Table 11: Logistic regression feature importance (in absolute values)

abs coef	period	event type
0.466722	2	SUBMIT PROBLEM
0.423006	2	SHOW ANSWER OF A PROBLEM
0.326207	1	CLICK NAVIGATION BAR
0.269215	5	SUBMIT PROBLEM
0.259197	3	PLAY VIDEO
0.242294	4	SHOW ANSWER OF A PROBLEM
0.218669	4	SUBMIT PROBLEM
0.208028	3	CLICK NAVIGATION BAR
0.147879	5	CLICK NAVIGATION BAR
0.146738	2	CLICK NAVIGATION BAR
0.135674	5	SHOW ANSWER OF A PROBLEM
0.134071	4	PLAY VIDEO
0.133285	1	PLAY VIDEO
0.077121	1	SHOW ANSWER OF A PROBLEM
0.075809	4	CLICK NAVIGATION BAR
0.073517	5	PLAY VIDEO
0.065158	3	SUBMIT PROBLEM
0.050152	3	SHOW ANSWER OF A PROBLEM
0.048661	2	PLAY VIDEO
0.006117	1	SUBMIT PROBLEM

results with other predictive studies done on the same BLE context. Notwithstanding some research done within BLEs are more in line with our study, such as the study done by [Zacharis 2015] that shows that there is predictive value of purely online data. Similarly, not all the events were found useful in the prediction with the use of binary logistic regression. Despite the fact that not all features used are comparable and that the study of [Zacharis 2015] did not include the dimension of time period-event connection, they showed similar results: two of the four decisive events registered by the student activity were "quiz efforts" and "files viewed", thus, the student interaction with the content had the highest value for prediction. The overall percentage of correct predictions by logistic regression was also comparable (0.84 LOOCV in our work versus 0.813 10-fold CV in [Zacharis 2015]). These similarities may indicate a high potential of logistic regression as a method for blended learning data predictive analysis. A second research done by [Nguyen et al. 2018] used linear regression over the different periods throughout the course, comparable to our aggregation of the data into periods, with an accuracy of 0.70. They show a comparable importance of the connection of a certain event with the period in which it has been recorded.

Since the focus of this study was on only one data set from one specific course it is not inconceivable that these models would perform differently on data from a different course. Furthermore, this study lacks a high enough amount of data points to independently verify the prediction model. For this reason a data set from another year is needed to verify the prediction model as proposed in this paper. This also means that the metrics in Table 7 should be taken carefully and that the LOOCV accuracy remains the most important metric to gauge the accuracy of the predictive models. Furthermore Table 6 shows that it is probable that the models over-fit when the number of features in the data set increases.

6 CONCLUSION AND FUTURE WORK

This study investigated the possibility of creating an accurate success prediction model in the context of a blended learning environment operating on edX Edge. We managed to create a model which can accurately (LOOCV accuracy = 0.84) predict student success in their final exam (fail or pass), despite the small sample size of the data set and the context of a BLE. As the predictive model is effective as early as from the middle of the entire course's period (including the exam study time), it provides an opportunity for early interventions to the educational process of at-risk students and facilitates the failure prevention. Additionally, the proposed model has given insights into the prediction process. It has shown that, for example, the importance of a certain event type is relative to the period in which it is registered (and vice versa), something that may be useful feedback to the teacher in terms of course design.

This paper has used only a small part of the potential information that resides in a raw data set collected from a VLE. There are still many potential areas open for further research. A more detailed time series analysis could be done by researching the volatility, or in this context the consistency in which a student would engage with the online course. Related to this, research done on for example stock markets, could be also applied on time series data from VLEs. Accordingly, session related features, such as "the time between sessions", "duration of a session", could be extracted as extra features from VLEs. Additionally, it could prove of great value to try and validate this model by using a data set from the same course, but from a different year, as a completely independent test set.

7 ACKNOWLEDGEMENTS

We would like to acknowledge the work of Cathlene Van Moorleghem and Isabel Tienza Perez while writing their master thesis "Exploring the data behind a conceptual modelling online course". The cleaned-up data set produced for this thesis was used in this paper.

REFERENCES

Antti Airola, Tapio Pahikkala, Willem Waegeman, Bernard De Baets, and Tapio Salakoski. 2009. A comparison of AUC estimators in small-sample studies. In *Proceedings of the third International Workshop on Machine Learning in Systems Biology (Proceedings of Machine Learning Research)*, Sašo Džeroski, Pierre Guerts, and Juho Rousu (Eds.), Vol. 8. PMLR, Ljubljana, Slovenia, 3–13. http://proceedings.mlr.press/v8/airola10a.html

Evandro B Costa, Baldoino Fonseca, Marcelo Almeida Santana, Fabrísia Ferreira de Araújo, and Joilson Rego. 2017. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior* 73 (2017), 247–256.

Shane Dawson, Liz Heathcote, and Gary Poole. 2010. Harnessing ICT potential: The adoption and analysis of ICT systems for enhancing the student learning experience. *International Journal of Educational Management* 24, 2 (2010), 116–128.

Daniele Di Mitri, Maren Scheffel, Hendrik Drachsler, Dirk Börner, Stefaan Ternier, and Marcus Specht. 2017. Learning Pulse: A Machine Learning Approach for Predicting Performance in Self-regulated Learning Using Multimodal Data. In *Proceedings of the Seventh International Learning Analytics &#38; Knowledge Conference (LAK '17)*. ACM, New York, NY, USA, 188–197. https://doi.org/10.1145/3027385.3027447

Fatima Harrak, François Bouchet, Vanda Luengo, and Pierre Gillois. 2018. Profiling Students from Their Questions in a Blended Learning Environment. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK '18)*. ACM, New York, NY, USA, 102–110. https://doi.org/10.1145/3170358.3170389

Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Trans. Inf. Syst.* 22, 1 (Jan. 2004), 5–53. https://doi.org/10.1145/963770.963772

Anne-Sophie Hoffait and Michael Schyns. 2017. Early detection of university students with potential difficulties. *Decision Support Systems* 101 (2017), 1–11.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 413–422.

Daniel Mueller and Stefan Strohmeier. 2011. Design characteristics of virtual learning environments: state of research. *Computers & Education* 57 (2011), 2505–2516. https://doi.org/10.1016/j.compedu.2011.06.017

Viet Anh Nguyen, Quang Bach Nguyen, and Vuong Thinh Nguyen. 2018. A Model to Forecast Learning Outcomes for Students in Blended Learning Courses Based On Learning Analytics. In *Proceedings of the 2Nd International Conference on E-Society, E-Education and E-Technology (ICSET 2018)*. ACM, New York, NY, USA, 35–41. https://doi.org/10.1145/3268808.3268827

Martin Oliver and Keith Trigwell. 2005. Can 'Blended Learning' Be Redeemed? *E-Learning and Digital Media* 2, 1 (2005), 17–26. https://doi.org/10.2304/elea.2005.2.1.17 arXiv:https://doi.org/10.2304/elea.2005.2.1.17

Manuela Paechter and Brigitte Maier. 2010. Online or face-to-face? Students' experiences and preferences in e-learning. *The internet and higher education* 13, 4 (2010), 292–297.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

Anthony G Picciano. 2014. Big data and learning analytics in blended learning environments: Benefits and concerns. *IJIMAI* 2, 7 (2014), 35–43.

María Jesüs Rodríguez-Triana, Luis P. Prieto, Alejandra Martínez-Monés, Juan I. Asensio-Pérez, and Yannis Dimitriadis. 2018. The Teacher in the Loop: Customizing Multimodal Learning Analytics for Blended Learning. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK '18)*. ACM, New York, NY, USA, 417–426. https://doi.org/10.1145/3170358.3170364

Yvan Saeys, Thomas Abeel, and Yves Van de Peer. 2008. Robust Feature Selection Using Ensemble Feature Selection Techniques. In *Machine Learning and Knowledge Discovery in Databases*, Walter Daelemans, Bart Goethals, and Katharina Morik (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 313–325.

Soham Sarkar and Anil K Ghosh. 2019. On perfect clustering of high dimension, low sample size data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP, 99 (2019), 1–1.

Stephen V. Stehman. 1997. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment* 62 (1997), 77–89. https://doi.org/10.1016/s0034-4257(97)00083-7

Roger Weber, Hans-Jörg Schek, and Stephen Blott. 1998. A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. In *Proceedings of the 24rd International Conference on Very Large Data Bases (VLDB '98)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 194–205. http://dl.acm.org/citation.cfm?id=645924.671192

I.H. Witten, E. Frank, and M.A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Science. https://books.google.nl/books?id=bDtLM8CODsQC

Nick Z Zacharis. 2015. A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education* 27 (2015), 44–53.