PREDICTING STUDENT SUCCESS: MACHINE LEARNING MODELS BASED ON ONLINE INTERACTION LOGS

W. Casteels, T. Perkisas

AP University of Applied Sciences and Arts Antwerp (BELGIUM)

Abstract

This paper explores the prediction of student performance characteristics based on their interactions with a learning management system. Leveraging log data from these interactions, we employ Machine Learning (ML) techniques to train and compare various models. Our primary focus is on predicting the final grades of students in a course, although the methodology is adaptable to forecast other performance metrics. The study encompasses the preprocessing of interaction logs, feature extraction, and the application of multiple ML algorithms. We compare the accuracy and efficiency of these models to determine the most effective approach for performance prediction. The results demonstrate that specific interaction patterns are significant indicators of student success, providing valuable insights for educators and administrators to enhance the online learning experience.

Keywords: Learning Management System, Machine Learning, student performance prediction.

1 INTRODUCTION

With the ubiquitous adoption of Learning Management Systems (LMS) in higher educational institutions comes an ever-increasing amount of interaction data, which enables the use of Learning Analytics (LA) [1]. LA provides insights into the learning behaviors of students and which patterns contained within the learning pathways might be indicative of academic success or failure [2].

The value of these interaction patterns lies in their ability to reveal underlying student behaviors on how they interact with the LMS that are often not immediately visible through traditional assessment methods. Beyond predicting the dropout rate for students [3], analyzing these patterns enables deeper insights into which interactions influence student outcomes, allowing for more personalized and timely interventions, not only after, but during the semester. For example, early identification of at-risk students through predictive models can enable targeted support, ultimately improving student retention and success rates. Although questions have been raised in the past on which data should be taken into account and what the best methodology is [4].

In this article several machine learning (ML) techniques are considered in order to assess their predictive potential on only the pure interaction data. No personal data (such as socio-economic status) pertaining to the students is taken into account. Besides the reduction in data depth, we also strive to keep an explainable component to the prediction, i.e. which interaction contributes the most to the prediction. This result gives valuable insight into differentiating behaviors and the possibility to improve how courses are presented to students, by eliminating the possibility to perform certain interactions. These results vary from course to course as each has unique learning pathways and varying degrees of online components.

The main guiding research questions are:

- Using only interaction data, what is the prediction potential of several ML techniques?
- Can the main contributors to the prediction be identified?

2 METHODOLOGY

Data consists of real anonymized interaction logs of several Moodle courses all originating from the same degree program. These courses vary in their online component; for some the offline component of the course grossly outweighs the online component, however, in all considered courses, the course material is to be found online. An overview of the courses in terms of their number of users, individual pages, and sessions. We define a session as consecutive interactions with no more than an hour in between, performed by the same student.

We present this data as input to the ML techniques in the form of the frequency of visits of specific pages and transitions of one page to the next per student. We also include the transition between pages as features, in order to see if certain transitions between the pages are more indicative of a certain result. Lastly, we also include the timestamps as features, aggregated as a frequency of activity at the level of weeks.

In this paper, we consider the following machine learning techniques: (a) logistic regression [5], (b) XGBoost [6], and (c) Neural network architectures represented by a recurrent neural network (RNN) [7]. The choice of these techniques is given by increasing complexity and decreasing interpretability. We will be using the sklearn and torch python package implementation for these techniques.

As this research aims to predict the final score students will attain for a certain course, the question of granularity rises, which influences the number of classes we consider in the classification problem. A classification into the integer value of the student's final score results in 21 classes (integer values 0 through 20). Alternatively, we could consider 3 classes (fail, pass, and distinction) or 2 (pass/fail). In what follows, we shall represent the results as a confusion matrix of the 21 classes problem. This allows a higher degree of interpretation (e.g., within how many points the prediction is accurate). Lower granularity is also easily aggregated from this matrix.

Furthermore, we consider two feature selection methods. The first consists of establishing a threshold for interactivity. While this might ignore rare events, it captures the most statistically represented pathways to results. This naïve feature reduction has the added benefit that it works on the one hot encoding representation without aggregated the different features. This helps us establish a baseline of the most interpretable way to implement both the representation of the data and feature reduction. The second representation (and feature reduction method) consists of Singular Value Decomposition (SVD) [8].

In order to not introduce changes in the optimalisation, in all the proposed machine learning algorithms, the cross-entropy loss function [9] is used for optimalisation.

The performance metric used is that of the F1 score [10]. This is, however, an aggregated score, calculated from the precision and recall, which are derived from the classifications True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) classifications. In the results section, we will always include examples of the confusion matrix, such that insight can be given into the performance of the system without aggregation. As such, insights can be gained into the range in which the model maps certain results: e.g., a model that maps results within 2 points of its true label performs objectively better than a system that maps them within 5.

Table 1. Overview of the parameters of all the considered courses.

	# users	**# sessions**	**# pages**
Course 01	62	2713	54
Course 02	36	3061	58
Course 03	64	3175	13
Course 04	74	2899	128
Course 05	45	2144	34
Course 06	58	4531	59
Course 07	55	1002	24
Course 08	45	799	39
Course 09	53	794	29
Course 10	51	1334	38
Course 11	43	2702	36
Course 12	45	702	11
Course 13	44	1192	49
Course 14	52	1857	46
Course 15	44	666	30
Course 16	49	1435	25
Course 17	45	5579	99
Course 18	55	1421	34
Course 19	50	3341	44

Course 20	45	1072	29
Course 21	45	1070	32
Course 22	41	1954	22
Course 23	61	1645	38
Course 24	75	3634	34
Course 25	43	429	16
Course 26	59	2480	61
Course 27	65	4260	39

3 RESULTS

For each machine learning technique, the confusion matrices of the courses that attained the lowest and highest F1 score are shown, together with the histogram of F1 scores of all the courses. In the confusion matrices, the vertical distance of the non-zero elements from the diagonal within the same column give the spread range within which the model predicts the result.

3.1 Logistic Regression (one hot encoding)

The multinomial logistic regression model chosen uses a limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS), chosen for both its limited use of computer memory and its robustness [11]. The results are shown in figure 1.

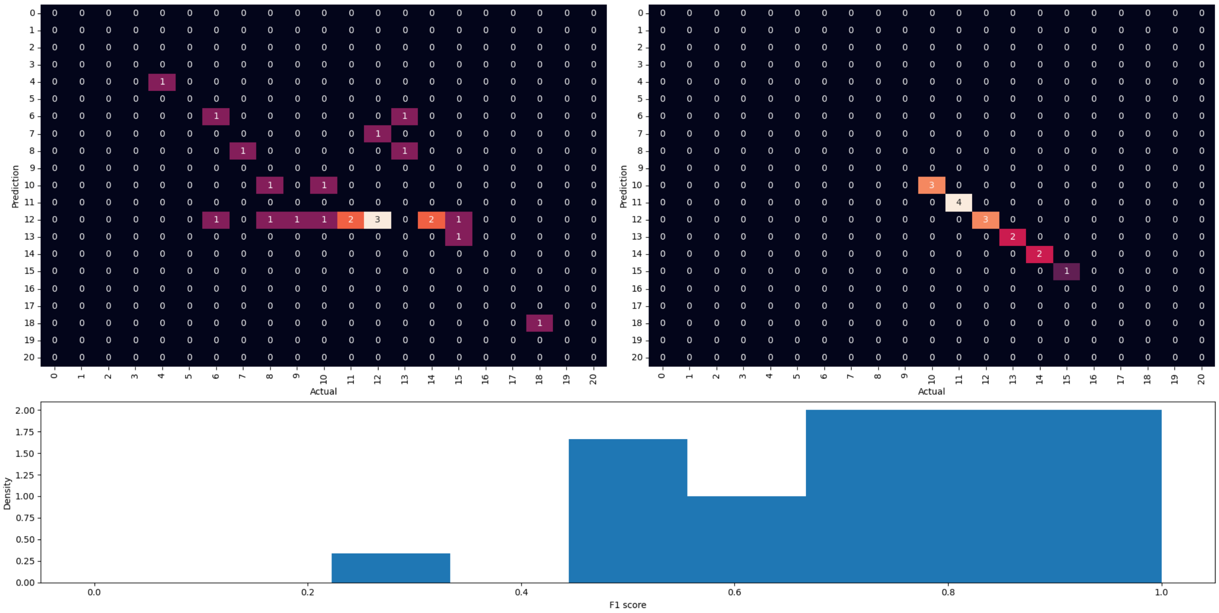

Figure 1. The confusion matrix for the worst scoring course (top left), best scoring course (top right) and the histogram of F1 scores for all courses (bottom) for the logistic regression with one hot encoding.

3.2 Logistic regression (SVD)

In order to study the effect of another representation and feature reduction, we run the same logistic regression model as before, with the extra preprocessing step of reducing the feature input by 25%.

The results are shown in figure 2.

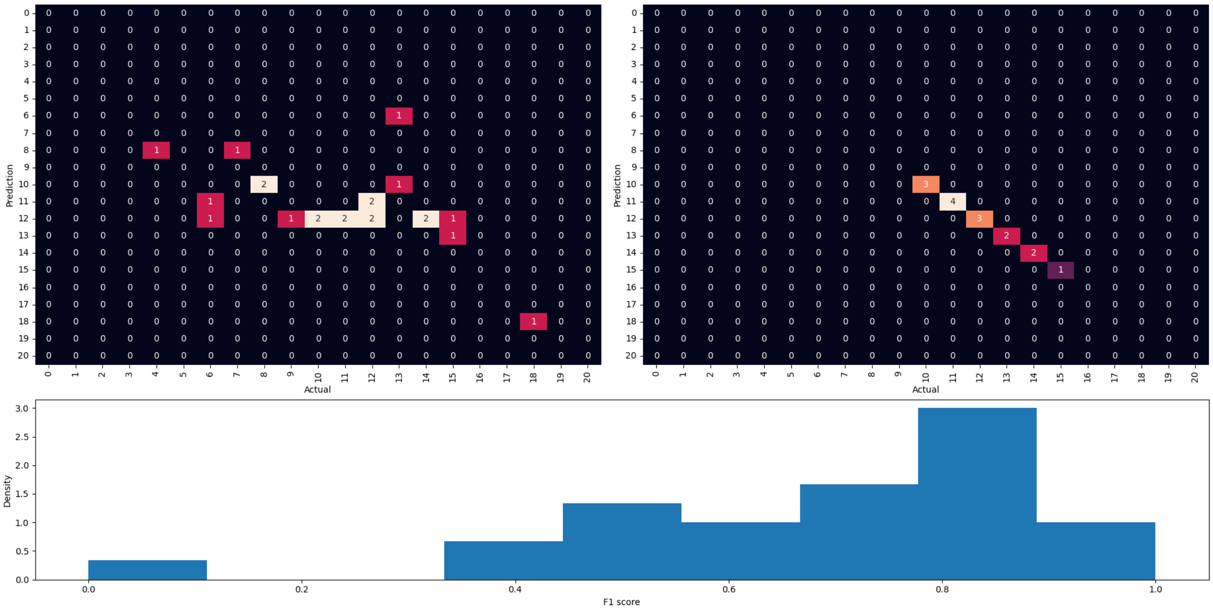

Figure 2. The confusion matrix for the worst scoring course (top left), best scoring course (top right) and the histogram of F1 scores for all courses (bottom) for the logistic regression model with SVD preprocessing.

3.3 XGBoost

For the XGBoost algorithm, a softmax for multi-class classification is used.

The results are shown in figure 3.

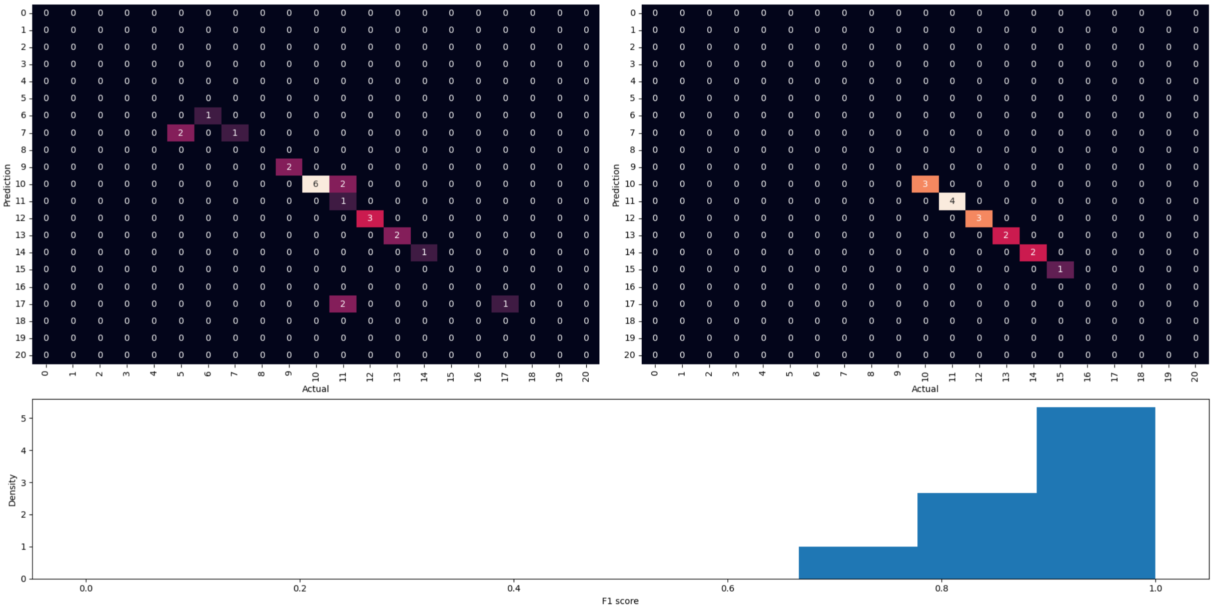

Figure 3. The confusion matrix for the worst scoring course (top left), best scoring course (top right) and the histogram of F1 scores for all courses (bottom) for the XGBoost model.

3.4 Neural network (LSTM architecture)

For the neural network model, we have chosen to use a LSTM architecture (64 units, trained for 30 epochs with a learning rate of 0.01).

The results are shown in figure 4.

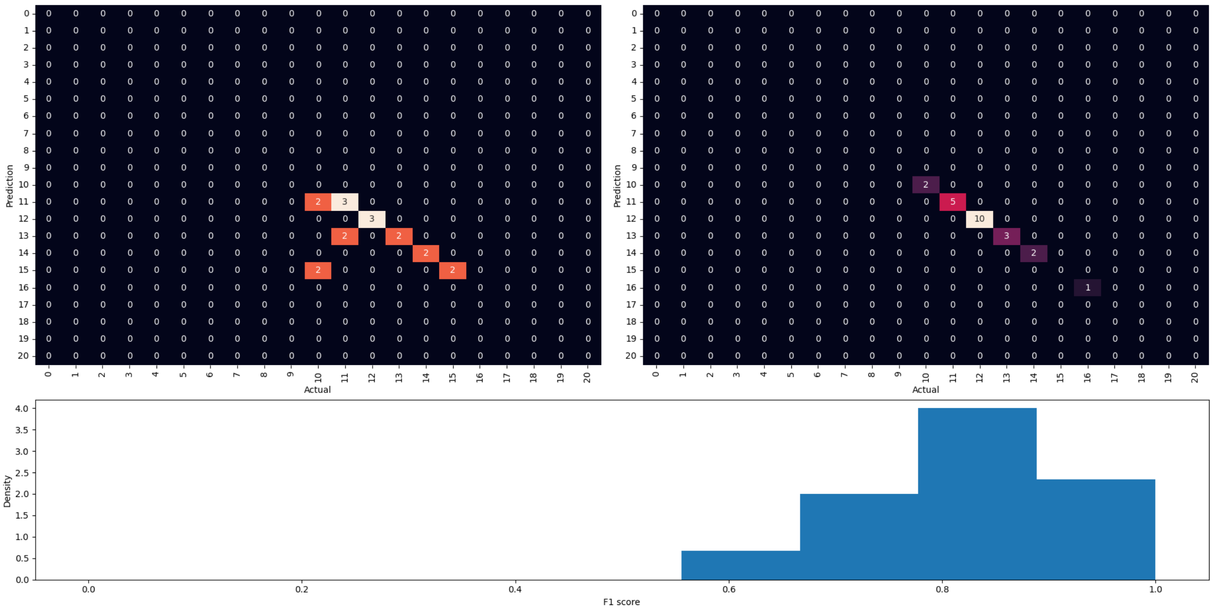

Figure 4. The confusion matrix for the worst scoring course (top left), best scoring course (top right) and the histogram of F1 scores for all courses (bottom) for the LSTM model.

4 DISCUSSION

We have shown the performance of 3 common machine learning techniques (logistic regression, XGBoost, and LSTM), deliberately chosen for their progressively higher complexity and data requirements. With this higher complexity comes a gain in performance; we can clearly see that the histogram of F1 scores shifts toward the upper boundary for XGBoost and the LSTM model. This is unsurprising as unlike logistic regression, XGBoost and the LSTM architecture do not assume linearity. The use of SVD for the logistic regression not only made the result less interpretable, but also failed to improve the performance. This is likely due to the combination of sparse data and its categorical nature. SVD was chosen in conjuncture with logistical regression as they both assume linearity. In future work, other techniques such as Principal Component Analysis (PCA) or Latent Dirichlet Allocation could be considered.

As this analysis is done on real life courses containing both an online and offline component, not all data is captured. For logistical regression, the model performs worst for course 12, and best for course 2. For XGBoost, the model performs worst for course 25 and best for course 4. For LSTM, the model performed worst on course 3 and best on course 5, but also on course 11. While for the model that assumes linearity there is an indication that less variability in the pages and therefore the possible learning pathways would cause a worse performance, it is too early to draw conclusions. A more systematic study of the role of course parameters on performance needs to be conducted over several degree programs. Likely, the distribution of results also plays a key role in the balance of classes in conjuncture with the number of sessions. It stands to reason that insufficient data and insufficient representation of classes are two of the main contributors towards this problem.

In terms of interpretability, the logistical regression gives access to the weights with which all inputs contribute to a decision. By including immediately consecutive interactions as well as the temporal information, the importance of these dimensions of the interaction could be proven. In all courses, these were present within the top 5 of highest contributors. When representing courses as a graph of nodes and edges (which represent learning pathways) the inclusion of the immediate consecutive interactions enables the insight into the importance of the edges. The inclusion of the temporal dimension gives insight into when a learning pathway becomes important.

One added difficulty we would like to draw attention to is the inconsistency of results, i.e., there are always students that with a minimal of effort (interactions) will score high, while others that put in considerable effort fail the course. This has to be taken into account in the interpretation of the false negatives and false positives respectively. Failing despite exhibiting right behavioral patterns and succeeding despite undesirable behavioral patterns: these aspects of reality that are not captured in the input data. This kind of type I and type II errors could potentially be indicated through the tracking of students over multiple courses, identifying such statistical outliers. However, this analysis is beyond the scope of this study.

Lastly, the models presented in this work show promise as on the entirety of a degree program, the confusion matrix tended towards diagonality, i.e. the off diagonals that contain non-zero elements in the confusion matrix tend to be of lower order. This can be interpreted that the models presented here are capable of predicting results within a reasonable range of the true result.

5 CONCLUSIONS

Our study has shown that the machine learning models considered in this work can be trained to predict student results, and this for fine granularity above binary pass/fail or pass/fail/distinction schemes. However, we have also indicated that all results need to be scrutinized as the effectiveness of these techniques can vary given the nature of the course. The inclusion of immediate consecutive patterns and the temporal information is a plus as they contribute significantly to the classification decision. Further study is needed to define the parameters for which a course can be said to be predictable and when these inputs become significant.

ACKNOWLEDGEMENTS

We acknowledge the Industrial Research Fund of the Flemish government (IOF) for funding this research.

REFERENCES

[1] L. Márquez, V. Henríquez, H. Chevreux, E. Scheihing, J. Guerra, “Adoption of learning analytics in higher education institutions: A systematic literature review,” *British Journal of Educational Technology*, vol. 55, issue 2, pp. 439-459, 2024.

[2] M. Shoaib, N. Sayed, J. Singh, J. Shafi, S. Khan, F. Ali, “AI student success predictor: Enhancing personalized learning in campus management systems,” *Computers in Human Behavior*, Volume 158, 108301, 2024.

[3] W. Casteels, L. Herrewijn, E. De Bruyne, “An example towards the responsible use of AI for higher education: explainable student dropout predictions”, *INTED2023 Proceedings*, pp. 4330-4336, 2023.

[4] A. Sønderlund, E. Hughes, J. Smith, “The efficacy of learning analytics interventions in higher education: A systematic review,” *British Journal of Educational Technology*, vol. 50, issue 5, pp. 2594-2618, 2019.

[5] C. Starbuck, " Logistic Regression" in *The Fundamentals of People Analytics: With Applications in R,* 223-238, Cham: Springer International Publishing, 2023.

[6] T. Chen, C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[7] A. Sherstinsky, “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network,” *Physica D: Nonlinear Phenomena*, Volume 404, 132306, 2020.

[8] L. N. Trefethen, D. Beau, *Numeric Linear Algebra*. Philadelphia: Society for Industrial and Applied Mathematics, 2022.

[9] A. Mao, M. Mohri, Y. Zhong, “Cross-entropy loss functions: Theoretical analysis and applications,” *ICML 2023*, 2023.

[10] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp 861–874, 2006.

[11] D. C. Liu, J. Nocedal, "On the Limited Memory Method for Large Scale Optimization," *Mathematical Programming B*, vol 45, no. 3, pp 503–528, 1989.