Predictive learning analytics in online education: A deeper understanding through explaining algorithmic errors

Journal Item

Computers and Education: Artificial Intelligence 3 (2022) 100108

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Predictive learning analytics in online education: A deeper understanding through explaining algorithmic errors

Martin Hlosta [a,b,*,1], Christothea Herodotou [b], Tina Papathoma [b,c], Anna Gillespie [b], Per Bergamin [a]

[a] The Swiss Distance University of Applied Sciences (Fernfachhochschule Schweiz – FFHS), Schinerstrasse 18, Brig, 3900, Switzerland
[b] The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
[c] CODE University of Applied Sciences, Lohmühlenstrasse 65, Berlin, 12435, Germany

ARTICLE INFO

Keywords:
Data science applications in education
Distance education and online learning
Post-secondary education

ABSTRACT

Existing Predictive Learning Analytics (PLA) systems utilising machine learning models show they can improve teacher practice and, at the same time, student outcomes. The accuracy, and related errors, of these systems can negatively influence their adoption. However, little effort has been made to investigate the errors made by the underlying models. This study focused on errors of models predicting students at risk of not submitting their assignments. We analysed two groups of error when the model was confident about the prediction: (a) students predicted to submit their assignment, yet they did not (False Negative; FN), and (b) students predicted not to submit their assignment yet they did (False Positive; FP). We followed the principles of thematic analysis to analyse interview data from 27 students whose predictions presented FN or FP errors. Findings revealed the significance of unexpected events occurring during studies that can affect students' behaviour and cannot be foreseen and accounted for in PLA, such as changes in family and work responsibilities, unexpected health issues and computer problems. Interview data helped identify new data sources, which could be integrated into predictions to mitigate some of the errors, such as study loan application information. Some other sources, e.g. capturing student knowledge at the start of the course, would require changes in the learning design of courses. Our insights showcase the importance of complimenting AI-based systems with human intelligence. In our case, these were both the interviewed students providing insights, as well potential users of these systems, e.g. teachers, who are aware of contextual factors, invisible to ML algorithms. We discuss the implications for improving predictions, learning design and teacher training in using PLA in their practice.

1. Introduction

The problem of correctly identifying students at risk of failing or not completing their studies has emerged as one of the most prevalent topics in Learning Analytics (LA), and an ongoing problem in education in general (Ochoa & Merceron, 2018). Over the years, this very problem has been examined in diverse learning contexts, including high schools (Knowles, 2015; Lakkaraju et al., 2015), face-to-face universities (Arnold & Pistilli, 2012), distance higher education institutions (Kuzilek et al., 2015) and Massive Open Online Courses (MOOCs) (Gardner, Yang, Baker, & Brooks, 2019). A commonly used strategy to tackle this problem is identifying at-risk students using machine learning models, or Predictive Learning Analytics (PLAs), followed by a subsequent support intervention targeting flagged students such as a phone call, text or email.

A vast number of published papers aims to achieve the highest prediction performance, measured by widely accepted machine learning metrics, e.g., accuracy, F1-Score or AUC (Area under the ROC curve). Within these papers, several learning algorithms are usually compared. One of the LA community trends stems from its cross-disciplinarity - many articles employ discoveries in machine learning to improve predictive performance. For example, several papers embraced Deep Learning methods, especially Deep Neural Networks focused on the time dimension (Fei & Yeung, 2015; Prenkaj, Distante, Faralli, & Velardi, 2021; Tang, Peterson, & Pardos, 2016); and used Deep Learning Encoders (Ding, Yang, Yeung & Pong., 2019; Klingler, Wampfler, Käser,

* Corresponding author. The Swiss Distance University of Applied Sciences (Fernfachhochschule Schweiz – FFHS), Schinerstrasse 18, Brig, 3900, Switzerland.
E-mail address: martin.hlosta@ffhs.ch (M. Hlosta).
1 Present address (Martin Hlosta): Fernfachhochschule Schweiz (FFHS), Schinerstrasse 18, Brig, 3900, Switzerland.

https://doi.org/10.1016/j.caeai.2022.100108
Received 8 August 2022; Received in revised form 25 October 2022; Accepted 14 November 2022
Available online 21 November 2022

Solenthaler, & Gross, 2017). Another piece of research examines specific problems associated with predictions, including the lack of legacy data (Ding, Wang, Hemberg & O'Reilley., 2019; Hlosta, Zdrahal, & Zendulka, 2017), the presence of fake student accounts (Alexandron, Yoo, Ruipérez-Valiente, Lee, & Pritchard, 2019), or the generalisability and transferability of the models in different contexts (Riestra-González, Paule-Ruíz, & Ortin, 2021).

Even in the case of a very accurate model, machine learning models produce errors (Springer & Whittaker, 2018), often attributed to irreducible errors (Blanzeisky & Cunningham, 2021). Such cases are not in all cases the result of wrong models but stem from the complexity of the modelled world and the inability to capture all the factors influencing the dependent variable. In a cross-disciplinary field such as LA, not having a perfectly accurate model, but one that is informed by a learning theory could still help to understand student learning by diagnosing factors that influence student success (Archer & Prinsloo, 2020). As Kitto, Shum, and Gibson (2018) argue, striving for the best performance metric should not be the goal of LA and having imperfect models does not necessarily mean these should not be deployed. On the other hand, error analysis can prevent putting a wrong, although accurate, model into production. For example, outside the LA domain, Ribeiro, Singh, and Guestrin (2016) demonstrate the wolf vs husky class example, which shows that, only after analysing the properties of the model, it becomes apparent that the husky is not identified by its traits but rather by the snow that surrounds it. The same husky in a forest without snow would immediately result in it being classified as a wolf. Such errors need to be corrected before deployment in the real world.

As LA becomes more mature, it is essential to reflect on and understand how predictive models behave and why underlying errors occur (Gardner et al., 2019; Taylor, Veeramachaneni, & O'Reilly, 2014). For example, Taylor et al. (2014) propose that error analysis should be presented in the rationale of a paper dealing with predictions. In education, as well as in other domains, research papers on fairness and model transparency have emerged. Such papers discuss model understanding and the errors they make (e.g., Kizilcec & Lee, 2021; De Laet, Millecamp, Broos, De Croon, & Verbert, 2020). However, as it will be shown in Section 2, model transparency focuses on model explanation rather than on targeting specific errors.

On the other hand, fairness quantifies the errors in relation to different groups, overlooking the bigger picture that these might not be the primary errors a model makes. Only a few studies focus directly on the errors that predictive models make (Lakkaraju et al., 2015; Qiu et al., 2016). A purely quantitative study might not be able to identify all underlying factors explaining errors, as many of such errors relate to activities not captured by the collected data ((Hlosta, Papathoma, & Herodotou, 2020); Schumacher & Ifenthaler, 2018).

The present study contributes to filling the literature gap by explaining errors in predictions while taking a qualitative research approach. The study involves 27 in-depth interviews with undergraduate students that models wrongly predicted as being/not at risk of failing their next assignment. In our analysis, we treated *false positive* (FP) and *false negative* (FN) errors separately, assuming that they would be explained by different underlying factors (Archer & Prinsloo, 2020). It is important to highlight that we follow the notation in Archer and Prinsloo (2020). We refer to FP as students who are predicted to be at-risk yet succeeded and FN as students who failed despite being predicted to succeed, and we are not using a reversed notation (e.g., Anderson, Boodhwani, & Baker, 2019). Our study aimed to answer the following research questions (RQ):

- RQ1: What factors explain errors in predictions as self-reported by 'false negative' students?
- RQ2: What factors explain errors in predictions as self-reported by 'false positive' students?

Section 2 reports existing research related to addressing errors in predictive learning analytics. Section 3 presents the methods and materials of this study. Results are presented in Section 4, followed by a Discussion in Section 5. The paper finishes with conclusions and future work in Section 6.

2. Acceptance, explainability and error related research in learning analytics

2.1. Model transparency and fairness

Analysing the errors is directly connected to *model transparency* (García-Martín & Lavesson, 2017) and *fairness analysis* and algorithmic bias (Baker & Hawn, 2021). **Model transparency** refers to the ability of a model to provide interpretability and explanation of its predictions (Mathrani, Susnjak, Ramaswami, & Barczak, 2021). Explaining which factors are used for automated decisions is important for several reasons, such as to satisfy legal requirements (Wachter, Mittelstadt, & Floridi, 2017). Black-box models relate to lower trust by users in predictive systems (Mathrani et al., 2021), which can hinder their adoption (Lakkaraju, Kamar, Caruana, & Leskovec, 2019; Wachter, Mittelstadt, & Russell, 2017). However, one of the challenges in the education domain is quantifying the added value of transparency to the stakeholders (De Laet et al., 2020). For example, Springer and Whittaker (2018) showed that providing explanations can negatively affect some users' trust when the explanation reveals algorithmic errors. In another study, both factual and counterfactual explanations were found to be inappropriate when the output did not meet what the user expected (Riveiro & Thill, 2021). As pointed out by (Dietvorst, Simmons, & Massey, 2015), users may not trust predictions when they notice that the algorithm makes errors. Just the fact that a model is transparent might not be enough, and (Springer & Whittaker, 2018) suggested algorithmic error presentation as one of the factors to focus on during designing user-facing predictive systems, especially when the cost of error is severe.

Contrary to model transparency, the work on **fairness and bias** directly focuses on the errors of the algorithms (Kizilcec & Lee, 2021). It investigates whether models predict the outcomes for all student groups of interest, usually gender, or ethnicity, with similar accuracy or generate a bias towards any group (Lee & Kizilcec, 2020). Predictions of under-represented minority ethnic students have lower overall accuracy, often attributed to having fewer students (Anderson et al., 2019; Yu, Li, Fischer, Doroudi, & Xu, 2020). However, it is important to break down predictive errors into false positives and false negatives, as they can have different significance depending on the prediction task. For example, for the task of predicting at-risk students, higher false alarms (false positives) on a minority ethnic group can be less harmful than not identifying them as at risk (Anderson et al., 2019; Bayer, Hlosta, & Fernandez, 2021). The first error might result in offering an intervention when not needed, which is not as problematic as not helping students when they need it (Bayer et al., 2021). The importance of errors reverses for prediction tasks where the positive outcome is more important for the future outcome, i.e. automatic admission system to a university. In this case, historical bias against minority students might deny them a place at a university (Kleinberg, Mullainathan, & Raghavan, 2017). Existing literature consistently reports that under-represented minority students are more often predicted with a negative outcome, e.g. low grade or not submitting the assignment, than the majority counterpart. This is true both for the negative outcome being the main predicted label (Bayer et al., 2021; Yu, Lee, & Kizilcec, 2021) as well as for the positive one (Anderson et al., 2019; Lee & Kizilcec, 2020; Yu et al., 2020). On the contrary, White students are less likely to be identified as at-risk than other groups (Anderson et al., 2019). In terms of getting closer to the source of errors, training only population specific models, i.e. only using minority students data, did not increase the accuracy of predictions on these populations (Anderson et al., 2019; Bayer et al., 2021). Therefore, it seems that under-represented students can benefit from a larger sample, even when the rest of the sample are White students.

Even though these studies quantify different types of errors according to different demographic groups, they do not consider contextual factors related to students (Yu et al., 2020). They showed that bias might come from institutional data, e.g. ethnicity, and gender. When these data sources were excluded, and only clicks were considered, errors diminished, yet lowered predictive accuracy. However, excluding protected attributes such as gender and race from predictions did not decrease the errors for these protected groups (Yu et al., 2021). This suggests that other factors might be behind these errors, and further examinations are needed to reveal them. Towards this direction, in this paper, we use a qualitative methodological approach to capture factors explaining false positives and false negatives as reported by students whose predictions presented errors.

2.2. *Studies analysing errors in predictions*

The current study builds on the methodology and early findings of the quantitative analysis proposed by (Hlosta, Zdrahal, Bayer, & Herodotou, 2020) and early results by (Hlosta, Papathoma, & Herodotou, 2020). The latter was a large-scale quantitative analysis identifying factors behind FP and FN prediction errors (FP = predict as not submit but did submit; FN = predict as submit but did not submit). This study focused on predicting assignment submission two weeks before the first assignment deadline in 62 courses across three years, producing 50,442 predictions for 36,352 students. Contrary to previous quantitative studies, FP and FN were treated separately, as manifesting different kinds of errors. Predictions were enhanced with future student activities, i.e. what happened in the subsequent two weeks following the predictions.

Moreover, as the predictions were course-specific, course context features such as the weighting of assignments and course length were considered. Then a decision tree model was built for FP and FN separately. For FP, the underlying decision tree was able to explain 50.91% of the FP errors with 75.29% precision (195 students). The strongest attribute proved to be the number of clicks one week before the first assignment deadline with a confidence of 0.77 (164/212 students with confident predictions submitted). Further filtering to students younger than 37 years increased the confidence to 0.82 (138/168 students) and for students with A Levels[2] or higher to 0.88 (101/115 students).

Contrary to FP, the model for FN only distinguished 18.73% of the errors with a precision of 68.73% (1036 students). The only significant pattern discovered was related to a dramatic decrease in students' activity in the last week before the first assignment submission.

Overall, both types of errors were associated with a change in student activity after the predictions were generated; either a drop for students that seemed on track or a steep increase for at-risk students. The presence of younger students in FP might be related to their higher tendency to procrastinate (Beutel, Klein, Aufenanger, Brähler, Dreier & Müller, 2016). On the other hand, the presence of more experienced students among false positives might be associated with self-regulatory strategies, e.g. minimising the time needed to study (Imhof, Bergamin, & McGarrity, 2021). However, these assumptions could not be observed in the data. Moreover, change patterns in the activity did not reveal any reasons explaining it. This issue shows that the insights one can gain from a quantitative analysis can be limited, and thus, there is a need for additional data collection directly from students, whose predictions present errors.

There is research that directly analysed errors in PLA, although this was not their primary research objective. In particular, the work of Lakkaraju et al. (2015) dealt with identifying students at risk of not graduating on time in the context of US high schools. As part of their study, they utilised an approach based on frequent patterns and association rule mining to describe the errors of classification models. The method first identified all the frequent patterns with 80% support, i.e. the ones that were covered by 80% of students. Then two patterns with the highest ratio of false predictions were selected. The most prevalent error was detected in students with high Grade Point Average (GPA) and high absence rates. However, the analysis did not distinguish between FP and FN as our study does.

Further, Qiu et al. (2016) focused on predicting assignment grades and certificate completion in MOOCs. They recognised three types of errors in PLA models: (1) unpredictable failing students (FN), (2) unpredictable "succeeders" (FP) and (3) swing cases defined by Qiu et al. (2016, p. 101) as 'those students whose scores are hovering around the minimum score'. Interviews with FP (2) revealed that a large number of those students were individuals who had already taken a similar course in the past. Despite treating FP and FN separately, the error analysis was only mentioned in one subsection, and the interviews were only conducted with the FP group with a brief mention of the results. Our study moves a step forward by also examining the FN group within a new context, that of online and distance learning higher education. Within this context, Calvert & Hilliam, 2019 investigated a single course at The Open University, UK, particularly on why some students passed even though their predictions at the start of the course indicated a low probability of passing. They conducted ten interviews with FP students who had 'succeeded against the odds' to offer advice to other students on how to succeed in the future. Among the top tips from successful students was to "start early", which required restructuring the delivery of the course by opening the course three months before it started. Calvert & Hilliam's study was qualitative and like Qiu et al., 2016 they focused on FP students only. Their predictions did not include student activity in the Virtual Learning Environment (VLE). Moreover, their analysis did not focus on the largest model errors but did include swing cases of students (Qiu et al., 2016). They selected their student sample based on a threshold where the prediction of passing the course was 0–75%.

3. Methodology

3.1. *Predictive model*

Students at The Open University are required to submit several graded Tutor Marked Assignments (TMA), as a prerequisite before the final exam. The underlying predictive model, which was used to predict which students are at risk of not submitting their next assignment, used four types of data:

1. *Static demographics* - Index of Multiple Deprivation (IMD) of the area where the student lives, ethnicity, and gender.
2. *Other static data known at the start of a course* - previous study results, study workload, i.e. the number of courses studied, and indicators of whether the course is repeated.
3. *VLE interactions* - weekly aggregated number of clicks per different activity types (e.g., forum, pdf resources, HTML content, homepage visit, tutorial attendance[3]), and weekly summary activity.
4. *Previous assignment scores* - only used for predicting the second assignment and onwards.

The previous presentation of the same course (i.e. previous run, the same course in the previous semester) is used to train each model. The predictions are updated weekly, which means that for each week and course a separate model is trained. From another perspective, each week as many models are trained as there are courses. This allows us to tailor the predictions for specific learning design of each course. It is also very different from having only one model for all courses - "One-size-fits-all"

[2] Advanced Level Qualifications - a UK secondary education qualification, typically an entry requirement to be admitted to a university.

[3] These activities correspond to the Moodle plugins used, which can be different for each course based on its design.

(Gašević, Dawson, Rogers & Gašević, 2016). Each week, all students from the previous presentations are used to train the model. The predictions in the current presentation are only computed for those who have not yet submitted. Only data collected until that week are used to train the model. This ensures alignment with the live data when we cannot know the clicks collected in the future weeks.

Gradient Boosted Machine (GBM) model from the R gbm package (Greenwell, Boehmke, & Cunningham, 2019) is trained. For each of them, the hyperparameters are optimised via grid-search for the highest ROC AUC, with the subsequent package procedure to select the optimum number of the underlying trees. The resulting model is used to generate the predictions and the confidence of the prediction in the interval [0; 1]. GBM has been used in the deployed system (see below) since 2018, based on the results of the comparison of several machine learning models in all undergraduate courses in 2018–2019, achieving the highest ROC AUC. We selected the predictions two weeks before the first assignment (TMA1) deadline. Hence, in our study all students in one course have only one prediction. Our aim was not to compare how predictions change in time, but rather in a time identified as important for interventions. Two weeks before the deadline, the predictions have high accuracy (AUC = 0.8897). Yet there is still time for a possible intervention to change students' behaviour, such as incentivising students to submit their assignments. This window of opportunity is noted by previous research between 2 and 4 weeks (Gašević, Dawson, Rogers, & Gasevic, 2016; Tempelaar, Rienties, & Giesbers, 2015). As a result, for the 62 analysed courses, for TMA1 two weeks before the deadline, 62 separate predictive models were generated for this study.

The resulting predictions are made available for all teachers via the web dashboard application, combining overall information for the student cohort that a teacher is assigned to and individual student prediction and its justification. As GBM does not provide the explanation directly, a model agnostic model - LIME (Ribeiro et al., 2016) is utilised to compute the five most contributing factors leading to the prediction decision. This is depicted in Fig. 1.

3.2. Extended data and confident predictions

For the quantitative analysis, data from all predictions were put into one table and extended by information that was not present in the prediction data:

- *course context* – the length of the course, the number of assignments in the course, the weight of the assignment towards the final grade and
- *future student activities* – data from the weeks following the predictions, which were unknown during the prediction's generation.

Predictive models provide estimates trying to resolve the uncertainty about the external world, often with the existence of cases where the prediction is close to more classes. In education, we often see "swing cases" where there are chances that a given student may eventually flip to a positive or negative outcome (Qiu et al., 2016). We aim to analyse only clear errors, and therefore, we excluded such predictions and selected only cases the model was confident with. We applied a threshold of 0.85 and selected only predictions with confidence ≥ 0.85 for both classes, i.e., Submit and Not Submit.

3.3. Method of data collection

Sample and process of data collection:

After gaining ethical approval for our study, an email was sent to 527 F P and FN students across 16 first-year STEM courses at an open and distance learning university, running when the data collection started in September 2019. We narrowed the focus on predictions of the first assignment submission (TMA1), as previous research showed that most drop-outs are likely to happen at that point in the lifecycle of a course (Hlosta et al., 2017; Walker-Gibbs, Ajjawi, Rowe, Skourdoumbis, Thomas, O'Shea et al., 2019). Thirty-eight students expressed interest in participating in the study (6.83%) however, nine of them withdrew prior to the interview. Interviewees were offered vouchers of £15 as a "thank you" gift for taking part in the study. Interviews were conducted online via Skype or telephone by Authors C and D and were audio-recorded. Audio files were transcribed by a professional service affiliated with the university where the study took place.

Interview questions (see Appendix A and B) were designed to address the RQs and lasted between 20 and 40 min. Participants were asked questions about their study patterns (i.e. working in the morning, at night, working space, home, library etc.) and about issues/problems while preparing the assignment (TMA1). Twenty-seven (n = 27) semi-structured interviews were conducted with self-selected (only certain students responded) and purposive sample of online undergraduate students who met the FP and FN criteria, i.e., they were predicted as confident Submit (n = 13) or Not Submit (n = 14) in their first TMA. Eight out of 14 F P students studied one or two additional courses alongside the course they were being predicted on by PLA. Three students (n = 3) studied two other courses, two of whom dropped out of one of the courses, and one continued with all three courses and passed them. The remaining five students (n = 5) studied two courses in total consecutively. Out of 13 FN students, two students (P16FN; P21FN) in the end submitted their assignment, however with a large delay (20 and 39 days) and without their extension being recorded. More group descriptive characteristics about both FP and FN students are described in Table 1.

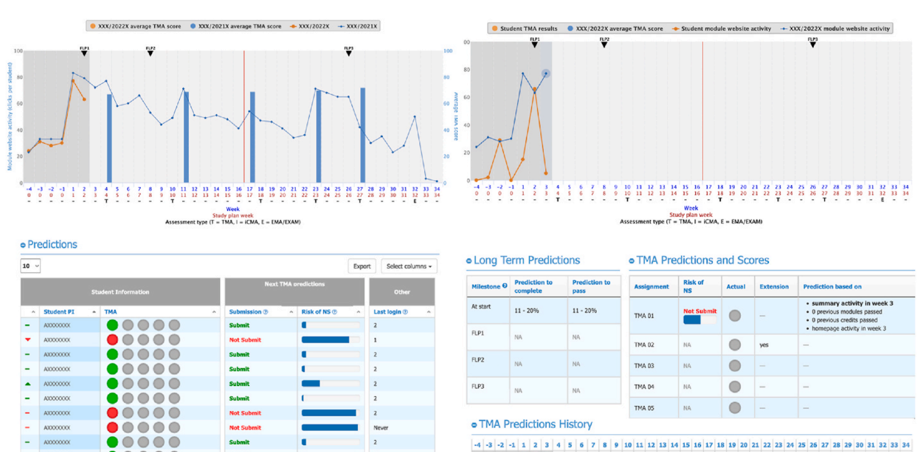

Fig. 1. Screenshot of the predictive analytics dashboard - (left) course overview (right) student detail view.

Table 1
Sample information for FP and FN.

	size	age_mean	age_std	age_med	CONT	NEW	F	M	Asian	Refused	White
FN	13	47.46	14.40	48	12	1	9	4	0	0	13
FP	14	31.86	8.33	32	10	4	6	8	3	1	10

3.4. Method of data analysis

We took a thematic analysis approach (Braun & Clarke, 2006) to analyse the interview data. Nvivo 11 software was used to develop/create codes and subsequent themes and to represent systematic patterns between different codes and subcodes across transcripts. Interview transcripts were grouped into two categories: (a) transcripts of students predicted to submit yet they did not submit (FN; n = 13) and (b) transcripts of students predicted not to submit yet they submitted (FP; n = 14). The interviews were coded and analysed by Authors C and D. To maximise the inter-rater reliability of coding, the first two transcripts were also coded by Author B. Coding differences were minimal and were extensively discussed in a meeting. An agreement as to which codes should be used was reached through discussion and negotiation. Codes were organised according to the student category they belonged to (FN or FP). The first interpretation of codes and production of a narrative to address the RQs was led by Author C. This was subsequently reviewed by Authors A and B and negotiated to reflect mutual understanding amongst co-authors from which the emerging themes were identified. Quotes used in the results section were anonymised using a unique number and the category a student belonged to (FN or FP). Table 2 presents the emerging themes and subthemes identified in the analysis across both student categories.

Table 2
Themes emerging from the thematic analysis of 27 student interviews.

Student group	Emerging themes	Subthemes
False Positive (FP) (n = 14): Students predicted not to submit, yet they did submit	Last Minute Catch Up	Studying last minute was explained by: - late finances - technical problems - family responsibilities - personality characteristics inhibiting study: low self-esteem, pride, - personality characteristics enabling last min study: self-organisation and motivation - tutor help - group work
	Previous Knowledge	- relevant professional experience - previous submission of assignment - correct "guessing" of answers
	External resources	- downloading material from VLE - studying from books - communication via e.g. WhatsApp (rather than VLE forums)
False Negative (FN) (n = 13): Students predicted to submit yet they did not submit	Finance-related problems	- delay in student loan and deregistration - delay in disability funding assessment - incorrect form completion
	Deferral for another course	- studying multiple courses at once
	Deferral due to family responsibilities	- family matters - there was no tutorial recording available
	Course design	- repetitive, boring - difficult content - low assignment weighting
	Tutor's unavailability	- assignment extension not granted
	Technical issues	- file zipping - internet access

4. Results

Results from the 27 interviews were clustered in two categories of students not correctly predicted by PLA. Section 4.1 presents findings from the 14 students predicted not to submit, yet they did submit (FP = 14), and Section 4.2 presents findings from the 13 students predicted to submit yet they did not submit (FN = 13). Fig. 2 Shows the number of clicks in the VLE of interviewed FP (left) and FN (right) students - two weeks, one week before the deadline, and one week after the deadline of the first assignment.

4.1. False positives (FP): students submitted despite being predicted as "not submit"

Three main themes emerged from the FP analysis: 1) Last Minute Catch Up, 2) Previous Knowledge, and 3) External resources. These themes are presented and discussed in the following paragraphs. Individual student predictions and their explanations are for comparison described in Table 3.

4.1.1. Last-minute catch-up

Low VLE activity triggered PLA to flag some students as at risk of not submitting their TMA. A range of factors explained students' low engagement, including a late study loan approval, technical issues, access to a computer, health problems, family issues, and low self-esteem. These issues resulted in students working last minute to catch up and submit their TMA. Despite the problems faced, these students managed to submit their TMAs on time. The patterns for these students from the quantitative analysis were related mostly to increased VLE activity in the last week before submission (see Fig. 2, left part). Also, most of these students had at least an 'A level' qualification.

In the quote that follows, one of the participants reports a delayed finance approval and the lack of technical equipment, which affected their motivation to study for the TMA (P04FP): *'My student finance was late to come through. So, I sat there and prepared not to start the year as I thought I was not going to get it in time …. I was (Abroad] as well, so I didn't have my laptop for some of the time"*. Once their finances were confirmed, they continued studying, yet they had to catch up and submit their assignment in a four-day period which was not how they would usually work. Another participant (P03FP) confirmed that their study pattern had been disrupted due to computer problems which they attributed to being the likely reason they were falsely predicted not to submit: *'I had to actually go to the library, and I had to book it in and I could only book it in for an hour a day because of the demand, so it kind of messed up how I work. It is so much easier for me when I am working in a comfortable environment, usually at home, and I am working on my own laptop.'* Another participant (P08FP) reported that they had difficulties due to family responsibilities: *'I have just put my 2-year-old to have his nap now, so that is my study time. I study when he is asleep … it is 12 p.m. to 2 p.m.-ish. Sometimes he only sleeps for one and half hours. It is unpredictable. And then I try from 8 p.m. to 12 midnight so 4 h in the evening'* (P08FP).

The same participant, P08FP, together with P27FP, were contacted by a tutor (P08FP; P27FP), who provided help on ways to catch up with the printed materials. P27FP, although they passed a test on assessing student's readiness at the start of the course, they presented themselves

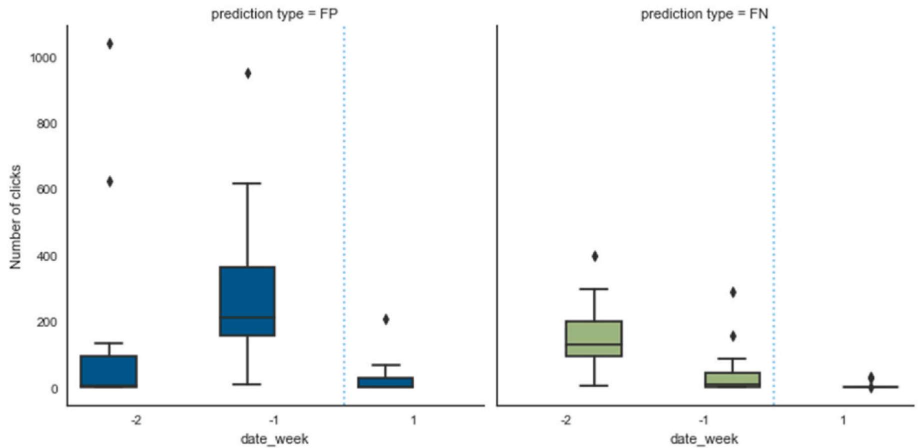

Fig. 2. VLE activity as the number of click of interviewed FP (left) and FN (right) students, two weeks, one week before the deadline, and one week after the deadline of the first assignment.

Table 3
Predictions and their explanations for FP by quantitative and qualitative analysis.

Participant	Risk NS	TMA score	Course Result	Clicks 1 week before and in TMA week	Prediction Explanation	Quant Explanation of wrong prediction	Participant Explanation of error
P07FP	95%	40–50	Pass	36 $\rightarrow$ 372	Education $<$ A Level Low VLE activity Low VLE activity before the course start	TMA week high activity	External resources, studying from books
P18FP	93%	$\geq$80	Fail	0 $\rightarrow$ 526	Repeating course No VLE activity Low VLE activity before the course start	TMA week high activity Education $\geq$ A Levels	Last min catch up, tutor help, "pride", previous knowledge, previous assignment submission
P23FP	93%	60–80	Pass	0 $\rightarrow$ 188	No VLE activity	TMA week high activity Age $<$37 Education $\geq$ A Levels	PreviousKnowledge, last min catch up
P09FP	92%	50–60	Fail	0 $\rightarrow$ 72	No VLE activity	Age $<$37	Previous knowledge, relevant professional experience, external resources, communication via apps
P01FP	91%	$\geq$80	Distinction	27 $\rightarrow$ 209	Education $<$ A Level Low VLE activity	TMA week high activity Age $<$37	Last min catch up, assignment extension
P02FP	91%	$\geq$80	Distinction	9 $\rightarrow$ 147	Low VLE activity Unknown occupation	TMA week high activity Age $<$37	External resources, downloading VLE material, communication via apps
P27FP	90%	50–60	Fail	0 $\rightarrow$ 110	No VLE activity	TMA week high activity Education $\geq$ A Levels	Last min catch up, tutor help, low self-esteem, external resources, studying from books
P06FP	90%	50–60	Pass	1 $\rightarrow$ 436	Low VLE activity High study workload Repeating course Education $<$ A level	TMA week high activity Education $\geq$ A Levels	Last minute catch up, group work
P04FP	89%	$\geq$80	Pass	0 $\rightarrow$ 0	No VLE activity	Age $<$37 Education $\geq$ A Levels	Last min catch up, late finances, self-organised
P05FP	89%	50–60	Fail	0 $\rightarrow$ 0	No VLE activity Repeating course	Age $<$37 Education $\geq$ A Levels	Previous knowledge, previous assignment submission
P24FP	89%	60–80	Pass	3 $\rightarrow$ 190	Low VLE activity Education $<$ A Level Repeating course	TMA week high activity	External resources, studying from books
P22FP	89%	$<$50	Fail	114 $\rightarrow$ 10	Low VLE activity	TMA week high activity Age $<$37 Education $\geq$ A Levels	Previous knowledge, correctly 'guessed' answers
P08FP	87%	50–60	Pass	24 $\rightarrow$ 368	Looking after family Low VLE activity Low VLE activity before course start	TMA week high activity Age $<$37 Education $\geq$ A Levels	Last min catch up, family responsibilities, tutor help
P03FP	85%	60–80	Pass	0 $\rightarrow$ 45	No VLE activity	Age $<$37 Education $\geq$ A Levels	Last min catch up, technical problems, self-organised

as having low self-esteem; they did not believe they had the necessary background knowledge to pass the course: *'I said to my tutor that I had serious concerns because I knew that my maths wasn't up for what was actually expected of me in this [course].'* This participant, despite completing the course, failed the final exam. Two other participants overcame study problems due to being self-organised and motivated (P03FP; P04FP). As explained: *'Studying from home is not an issue for me because I am quite motivated to finish my [courses] […] I am usually good under pressure. I think it was a change of scenery or being surrounded by people, that is why it doesn't have the same result that it usually did. I think I did procrastinate a little, but I try not to, I always give myself an extra day just in case I have missed something, or I need to change something.'* (P03FP).

Another participant did not want to contact the tutor due to perceptions of 'pride' and because they had managed to complete their previous TMA. They reported: *'I think the reason I didn't [contact my tutor] was, half pride I guess, and two, I think, I got the TMA done, and I think I'm gonna get the other, next TMA done. I don't think it's gonna affect me, but, the tutors seem like they're very nice and they could help out … '* (P18FP). Despite the participant's perception of being on track after TMA1, they later failed the course.

4.1.2. Previous knowledge

Some participants stated that they already had the knowledge required for the first TMA, which meant that they did not need to study the VLE materials or would use them less often. Most of these students also had at least an 'A levels' qualification. One participant (P09FP) mentioned previous knowledge related to their experience: *'I am a designer already, so like from 3 years and I was expecting something a lot [more artistically] simulating.'* For two participants (P05FP; P18FP), their prior knowledge was explained by repeating the course they had failed in the past. As explained by one of them (P05FP): *'I just went through the assessment … most of it I have kind of done before but when I needed to do specific pieces of work, like pieces of study, I just kind of found what I needed.'* Interestingly, these two participants submitted their first TMA and reached course completion, yet they failed the course in the end. Such insights suggest that sole assignment submission does not suffice for a student to pass a course; extra support is needed by, for example, a tutor to ensure a student submits high-quality work that can help them pass a course.

Another participant who faced health issues and could not prepare for the TMA 'guessed' correctly TMA answers due to prior knowledge of the topic. As explained: *'So, basically what I did, I just, where I could I guessed. And I think some of the maths I already knew.'* (P22FP).

4.1.3. External resources

Some FP participants mainly studied offline or used external VLE materials to study. PLA considered these students as not engaged and at risk due to their limited VLE engagement. The increased activity observed after the predictions were produced was explained by using the VLE at a later time to prepare and submit a TMA: *'I studied from the books, and then I only used the websites when preparing my TMAs.'* (P24FP). Offline study was the preferred mode of study for some participants, especially those required to travel: *'The biggest thing is that I downloaded the materials, so I am doing it offline. I am not interacting with the website as much as other people.'* (P02FP).

Another participant explained their preference to use external resources such as audio materials as English is not their first language: '*I was more attached to the reference books instead of the course book cos I found the language pattern and understandability much easier than that of the prescribed course book … English is not my first language. So brainstorming with voice scanning does help, they keep me on track and fast*' (P07FP). Two participants (P02FP; P09FP) used WhatsApp to communicate with other students rather than using the course forum. WhatsApp was used because participants would receive direct responses compared to slow replies in the forum: *'I just use WhatsApp because it is more instant if I have a question'* (P09FP). In some student cases, a combination of factors was observed to explain wrong predictions, such as external resources and previous knowledge, see previous quotes of (P09FP), and external resources with last-minute studying: *'And I just end up forgetting that there's tutorials on and that I can't really participate at my home. So, I'm missing out there and the material as well, because it's all online, it's quite awkward, so after speaking with my tutor, I think it might have been after that one actually. She said about having the materials actually posted out to me. So, then I just end up having them sent out to me so I can actually do exactly the same work but in paper-based instead.'* (P27FP).

4.2. False negatives (FN): students do not submit despite being predicted as "submit"

Seven factors explained why participants did not submit their assignment despite being predicted as "Submit" (n = 13): (1) Finance-related problems, (2) Deferral for another course, (3) Deferral due to family responsibilities, (4) Assignment extension, (5) Assignment weighting, (6) Technical issues, and (7) Course Design. Apart from two participants, the rest of them were predicted as submitting mainly due to high VLE activity in the prediction week. P12FN was due to high activity in previous weeks, and P14FN was due to the high quiz score, despite low VLE activity. Most FN could be characterised by an immediate drop in VLE activity in the week following the predictions. Individual student predictions and their explanations are for comparison described in Table 4.

4.2.1. Finance-related problems

Some of the FN participants reported issues with their student finances that inhibited a timely TMA submission. As explained by one participant, they had not received their student loan on time and were subsequently deregistered from the course: *'on Dec* 12th *I was deregistered by the (university), and on Dec* 17th *Student Finance England wrote to me and said we will pay your course tuition* fees.*'* (P12FN). When asked if they approached the university to re-register, they reported that they did not think that would be possible; therefore, they did not pursue it. Another issue reported was about difficulties in getting assessed for disability funding: *'I was eligible to disable allowance and that I will go to* The Open University *to speak to the appropriate people to see what systems I [require, which] in my own case, it was a laptop so I could work in my own place. But there [were] no contact details on that […] I called 3 or 4 different numbers, and nobody seemed to know what I was talking about'* (P14FN). The same participant stated that they did not think they could get their finances in place on time for the cut-off date; thus, they decided to defer. As a high achiever, this participant continued to appear active on the VLE until the transfer to their other course was completed. Another participant explained that the loan was delayed because they had filled an incorrect form when applying for it: *'My student finance in August I filled the wrong form in, it's all my fault, and they sent a letter, but I never received the letter, so, I didn't know anything until the [institution], sent me an email … '* (P19FN).

4.2.2. Deferral for another course

Other participants chose to defer their registration until a later intake due to studying multiple courses. They stated that studying two courses while working and having family commitments became unmanageable, particularly as they also found the course content difficult. As explained: *'I just need to take the stress off me a bit, and I was going to defer to February and take it again then to space them out but what actually* happened *is I have got a lot on at work with my promotion stuff … year I am planning to do the September/February start'* (P11FN). This participant remained registered on their other course and reported that before submitting the first assignment, they had expected to continue studying both courses; hence they engaged in all the course activities online. Another student stated that they realised early on that they could not manage to study two

Table 4
Predictions and their explanations for FN by quantitative and qualitative analysis.

Participant	Risk NS	TMA score	Course Result	Clicks 1 week before and in TMA week	Pred Explanation	Quant Explanation	Explanation
P10FN	12%		Pass	59 → 45	High VLE activity Low workload High previous pass rate	–	Repetitive, boring course
P17FN	12%		Fail	320 → 6	High VLE activity Low study workload	Drop in activity Course with high VLE activity	Low assignment weighting
P26FN	11%		Withdrawn (–)	262 → 0	High VLE activity High previous pass rate	Low VLE activity and drop in TMA week	Assignment extension not granted
P15FN	8%		Withdrawn (reenrolled)	80 → 0	High VLE activity Low study workload	Low VLE activity and drop in TMA week Course with high VLE activity	Course design, difficult content
P11FN	4%		Withdrawn (reenrolled)	400 → 0	High VLE activity University Education Low Study Workload High score quiz	Low VLE activity and drop in TMA week Course with high VLE activity	Deferral for another course, studying more than one course and family responsibilities
P16FN	4%	≥80	Pass	67 → 12	High VLE activity Low study workload Full-time employed	Course with high VLE activity	Technical issues, internet access
P19FN	4%		Withdrawn (reenrolled)	97 → 85	High VLE activity	Course with high VLE activity	Finance-related problems, incorrect form completion
P14FN	3%		Withdrawn (reenrolled)	21 → 0	New Student High score quiz	–	Finance-related problems, delay in disability funding
P12FN	2%		Withdrawn (–)	4 → 3	High previous weeks VLE activity Low study workload	–	Finance-related problems, study loan and deregistration
P20FN	2%		Withdrawn (–)	155 → 236	High VLE activity	Course with high VLE activity	Technical issues, zipping files
P21FN	2%	60–80	Withdrawn (reenrolled)	319 → 94	High VLE activity	Drop in activity Course with high VLE activity	Deferral due to family responsibilities, no tutorial recording available
P13FN	2%		Withdrawn (reenrolled)	92 → 0	High VLE activity Low Study Workload High score quiz	Low VLE activity and drop in TMA week Course with high VLE activity	Deferral for another course, studying more than one course and job responsibilities
P25FN	1%		Withdrawn (reenrolled)	165 → 0	High VLE activity Low study workload	Low VLE activity and drop in TMA week	Deferral for another course, deferral due to family responsibilities

courses simultaneously. Hence, they decided to defer until a later intake: *'It is just that I realised very quickly that I had to drop one of them and [that is] the only reason that I dropped [this course]'* (P13FN). This decision was explained by starting a new job and the need to get used to a new set of responsibilities.

4.2.3. Deferral due to family responsibilities

A participant reported that their study at the university was affected by family responsibilities, and though they were studying one course at a time, they decided to withdraw. As reported: *'I was out of the country a lot because of family and legal matters, so I wasn't involved in* The Open University *stuff at all, quite often for weeks and months at a time, so I was permanently in a position of having to catch up'* (P21FN). This participant submitted the TMA with a delay but then withdrew, possibly because of a lack of online tutorials: *'I wasn't in the country to attend the physical tutorial, the face to face tutorial, and I haven't attended any online tutorials. And one of problems I have is [that] it appears, unless I'm missing it, there are no recorded tutorials for me to look at to catch up.'*

4.2.4. Course design

Two participants explained the reasons why they did not submit their assignments for reasons related to course design. One participant commented that they found the course repetitive and boring. This made them feel demotivated, and they decided to drop out of the course (P10FN). Another participant reported that they found the course particularly difficult as it has been a long time since the last time they studied. In addition, the possibility of failing affected their decision to not submit: *'I got as far as question 4, but I don't know whether my answers before then were correct anyway […] I didn't want to sort of fail and think oh I can't do this, I will give up now. Which is sometimes what I can be like, I want to stop before it [gets] to that point and give myself a better chance at a better grade'* (P15FN). Assignment weighting was another dimension discussed. One student explained the non-submission of the first assignment with reasons related to the 'weighting' of the assignment. As explained: *'I did see that [it's only 7% of overall coursework] actually […] and I just thought, 'it doesn't matter'* (P17FN). The student did not feel it was worth submitting as it would not affect their final grade much. Such insights suggest that assignment weighting should be reconsidered and potentially increased, making it more meaningful for students to complete it.

4.2.5. Tutor's unavailability

One of the participants tried to reach their tutor to request an extension. Yet, they did not receive a reply, more likely due to the tutor participating in a planned industrial action taking place at the time the student tried to reach the tutor. As explained: *'I emailed my tutor and said, "This is the situation, may I have more time?" I didn't get a response. It might have been because it was December or because I know some of the tutors were* *on strike'* (P26FN). This suggests that students should be aware of occasions when their tutor is not available (due to industrial action or another reason) and be offered alternative contacts for support. A no reply poses a risk to students' progress and engagement with their studies. For example, course chairs might consider asking students to contact the Student Support Team if their tutor is temporarily not available.

4.2.6. Technical issues

Another participant struggled with technical issues related to the submission, including the process of zipping files. As explained *' … but some of the requirements of especially of physics and space was interacting a great deal with software,..and … the technical side of for example zipping files and, you know, and when I set myself I think 15 h ….*so *that was more to do with my, my lack of skills than it was to do with the actual running of the course' (P*20FN*)*. Such issues suggest that live technical assistance should be available to support students, or students should be signposted to where they can find written help and guidance via the VLE. The same participant stated additional difficulties such as the language barrier as English was not their first language, illness preventing on-time submission, and a late request for an extension which was not granted. Despite these difficulties and not submitting the TMA, they remained registered and passed the course. P16FN had issues accessing the internet: *'I asked the tutor for an extension, […] normally when I'm away I can have access to the internet but unfortunately, this time I was stuck somewhere where there was no internet..'* The extension allowed the student to submit later and eventually pass the course.

5. Discussion

Despite a high number of research published on predictive modelling in Learning Analytics, particularly for how to build predictive models, very few studies have examined the errors these models produce. This is a particularly timely issue given the recently highlighted importance of increasing trust and transparency of such models. Our study aims to contribute to this gap by investigating errors in an online learning context of distance higher education. Based on the classification of errors by Qiu et al. (2016), our study focused on errors with confident prediction yet different outcomes, and both false negative (FN) (noted by Qiu et al. (2016) as Unpredictable negative cases) and false positive (FP) (Unpredictable positive cases). To understand the reasons explaining these errors, we interviewed and analysed data from 27 students whose predictions presented errors. In the next paragraphs, we discuss these factors drawing from relevant literature.

5.1.2. False positives

Three overarching factors were shown to explain errors in false positives (FP), that is, errors in the predictions of students who were showing NOT to submit their assignment, yet they managed to submit it: a) Students catching up on their studies last minute, b) prior student knowledge about the topic of the assignment and c) external resources such as studying offline. Students' prior or existing knowledge about the topic of an assignment aligns well with interviews by (Qiu et al., 2016), with confident FP showing the presence of skilful students who took similar courses already, thus not needing to study as much as other students. In both cases, these students are found to be much less active in the VLE, and these data probably trigger at-risk predictions.

The two other existing papers analysing the errors in learning analytics (Calvert & Hilliam, 2019; Lakkaraju et al., 2015) focused more on the swing cases, i.e. borderline students. Most likely, for this reason, the factors do not align completely with our study. Lakkaraju et al. (2015) showed that the errors both for FP and FN were mostly for students with positive and negative factors appearing together, i.e. high GPA & high absence or low GPA & low absence. These contrasting factors are likely to confuse the classifier and increase its uncertainty. This seems similar to FP errors explained by Previous Knowledge, where the low engagement resembles high absence. However, while GPA is present in the input data, in our case, the high level of previous knowledge is unobserved, which tips the scales towards a confident at-risk prediction.

The main themes that emerged from the analysis of borderline FP by Calvert and Hilliam (2019) were good organisation related to starting studying early and willingness to try different approaches to learning. Two FP students mentioned self-organisation in our study (P03FP; P04FP). In the case of Calvert and Hilliam (2019), the good organisation compensates for different student conditions at the start of the course. In our case, it was more recovering after difficulties faced during the course resulting in the last-minute catching up. We have not directly observed the theme of *willingness to try different approaches* as a result of escaping being at-risk. Still, some students had to adjust their study patterns to their conditions, e.g. studying offline due to travelling (P24FP). It is also important to highlight that predictions by Calvert and Hilliam (2019) did not utilise any VLE activity and were based on static indicators at the start of the course, while in our case, most of the predictions are later disproved by a change to the previous engagement in the VLE.

External factors referring to students studying offline is a theme mentioned in existing Learning Analytics studies as a factor limiting our interpretation of predictions (Baker et al., 2020; Rets, Herodotou, Bayer, Hlosta, & Rienties, 2021; Schumacher & Ifenthaler, 2018). When there is a choice, students may prefer to study offline from printed material rather than online using the VLE (Schumacher & Ifenthaler, 2018). This choice might sometimes be impacted by the learning design of a course, especially when course activities are not tied to the VLE. In cases when most of the activities are outside VLE, students also consider Learning Analytics data as less useful (Verbert et al., 2014).

5.1.3. False negatives

A different set of factors was shown to explain errors in the predictions of false negatives (FN); that is, students predicted to submit their assignment, yet they did not submit it. These factors included a) finance problems, b) deferral for another course or due to family responsibilities, c) the course design, d) unavailability of a tutor to provide an extension, and e) technical issues. Contrary to FP, there is limited research on errors related to FN. This might be due to a greater interest to investigate and capture the strategies students use to escape from their at-risk status and underlying issues causing difficulties than understanding factors behind students on track of their studies. Moreover, our previous quantitative study (Hlosta, Papathoma, & Herodotou, 2020) showed that the FN errors are more difficult to explain (18.73% FN vs 50.91% FP), possibly attributable to unexpected events. Qiu et al. (2016) hypothesised that the unpredictable nature of FN might be related to personal reasons, e.g. scheduling conflicts. This is consistent with the themes that emerged in our analysis, mainly deferral due to family responsibilities, tutor's unavailability and to some extent also finance. The first two events are challenging to predict as they are not recorded. In some cases, tutors will maintain a relationship with the student. They might know about their personal, family or health situation and provide pastoral support for students to help them through their difficult period and avoid dropping out (Herodotou, Maguire, McDowell, & Hlosta, 2021).

For the finance theme, the complication for the prediction is caused by both not considering the application for such loan in the models but also because of lack of knowledge such as what happens in the external loan companies that evaluate a loan. Student loans are associated with high stress and the risk of dropping out of college (Britt, Ammerman, Barrett, & Jones, 2017). Some studies also included the loan as an attribute of the predictive model (Delen, 2011). However, in our study, the issue for the predictions originates from not knowing the loan status. In the meantime, the student studies, and if the loan is approved late, the student does not submit the assignment and has to defer. This topic is, to our knowledge, not covered by the existing learning analytics literature.

As students needing loans are expected to come from low socio-economical backgrounds (De Gayardon, Callender, & Green, 2019), these issues are likely to contribute to the existing gaps between poor and affluent students (Chmielewski, 2019).

On the other hand, students in the theme 'Course design' reported reasons related to how certain courses have been designed by the University, including the weighing of an assignment or having difficult and boring material. These aspects affected their motivation to study and led to a drop in engagement and a non-submitting of an assignment. This was both due to internal motivation (boredom) and external (low weight of the assignment on the score). This is aligned with theories where motivation influences student persistence (Tinto, 1975) and expectancy-value theory (Wigfield & Eccles, 2002, pp. 91–120). The perceived value of these students decreased below a threshold that influenced their continuation.

5.2. Implications

Quantitative methods are rather dominant in the LA field, especially due to rich VLE traces. On the other hand, by interviewing students and using a qualitative methodological approach, our findings can fill the gap of unexplained variance from prior quantitative research. To gain insights into what is happening outside the VLE, we need to zoom out of the captured VLE data. Rather than hypothesising about possible reasons for the error and limitations of the VLE data, our study provides evidence of some of the reasons. For example, the study loan theme is rarely mentioned as a complementary issue for students that might otherwise be on track with their studies. It highlights a possible gap in identifying who of the students might be at risk of not getting a loan from external companies that can inform timely support.

Our study has implications related to three areas: 1) enhancing predictive systems, 2) updating teachers' training when using these systems and 3) informing course designers.

5.2.1. Improving predictive systems

The analysis of the interviews assisted in identifying additional data that can improve the predictions and prevent errors. This is crucial, especially for the FN errors, which can help identify students who would have failed without being noticed in advance. We stressed, in particular, the importance of finance and study loans. Student data can include a flag indicating whether a student applied for a loan and a status for the application if known. Despite the student being active in the VLE, not obtaining a student loan may mean that they are at risk of deferral. This could allow tutors to identify that the student does not need study help but rather financial advice.

Most PLA studies only consider data from a single course that a student studies, while they could include data about a student's overall workload. The interviews revealed that some students handle their studies well despite being enrolled in more courses, while others do not. Therefore, the predictions could use detailed VLE traces from other studied courses. Knowing that students are active and submit their assignments in all their courses may suggest that their workload is not an issue. This knowledge can resolve some of the potential FP.

Several errors resulted from a lack of information about the student's existing knowledge of a topic. Current data about previous student education and results, while helpful, in our study was shown not to be adequate. Also, for new students, previous results are not available. This would require capturing additional diagnostic data at the beginning of the course, which would require a change in the course design (see 5.2.3). Another avenue explored in research is collecting data from various sensors, such as EEG or skin galvanic response. For example, galvanic skin response might capture students' affective state (Bannert, Molenaar, Azevedo, Järvelä, & Gašević, 2017). For example, detecting bored students can help identify future disengagement. Such applications might be challenging in online and distance education as they would rely entirely on students using such devices.

5.2.2. Teachers' training

The fact that some predictions might be disproved should not discourage teachers from using PLA systems. Predictions should not be viewed and interpreted as determining a future result, i.e. the self-fulfilling prophecy (Rosenthal & Jacobson, 1968). Instead, they should be seen as an estimate of an outcome which would have occurred if no change occurred. The importance of a teacher to support students and provide pastoral care becomes crucial, as it can change a prediction and benefit students. In this respect, teachers should be offered training to gain insights into the benefits and limitations of PLA so that they develop the skills and knowledge needed to evaluate prediction outcomes and not discard them just for the mere fact that, in some cases, they may produce errors. Moreover, errors for highly confident predictions, despite not being frequent, will still occur in PLA. This is due to both inability to capture some information and the occurrence of unpredictable events in students' lives. Students might study outside the VLE due to their preferences in reading printed materials. A possible solution is to capture student preferences and plans prior to the start of a course.

Implications differ between FP and FN students. Information from FP suggests strategies that could help students escape potential dropouts. For example tutors could direct students to a recording of a tutorial and advise students to download it if they were not able to attend it online. Also, students could be instructed about 'emergency strategies' and time management practices to manage their studies. On the other hand, if a teacher knows about a student's preference to study offline, e.g. printed materials, they can consider this if such student is flagged as at risk due to low VLE engagement. Information about study practices and preferences could be communicated in an introductory email to students.

On the other hand, tutors should be informed about errors when students are predicted to submit confidently, including the instances where students study more than one course and may decide to defer a course in favour of another one; instances where students defer the course because their loan did not get through; instances where an unexpected event happens in the personal lives of students, or a student realises that their family responsibilities are too high and must defer the course. Although study workload is already highlighted to students, it is reasonable to remind them that a high workload can lead to potential problems in the future and encourage them to examine what their options are and the commitment they need to make to study and complete a course.

Another topic for teachers' training should include guidelines in case a tutor is unavailable, due to strike or for any other reason. This will help students be aware that the tutor may not respond. Potential training could make use of specific examples from the current study. Students might be informed in a guideline about contacting the Student Support Team in case their tutor is not answering.

5.2.3. Learning designers

In addition to providing personal support to students during a course, learning designers can remove some barriers and make a learning environment more inclusive. They should consider providing an automatic recording of tutorial sessions. This will help students who do not feel comfortable joining live sessions or students who cannot participate, e.g. due to travelling. Another issue to consider is reassessment weighting. While a high weight on early assignments might put pressure on students at the start of a course, a low weight might discourage some of them from working towards the assignment and submitting it.

Like teachers, course designers should also consider some diagnostic tests to capture students' prior knowledge. An assessment of whether a student is ready for a course is currently available in some courses. Yet, a review is needed as there was an instant with a participant noting that the test was not challenging and did not reflect the content of the course. Further, a test assessing digital literacy will prevent situations in which students will be reporting that they do not know how to zip a file of their assessment TMA for a submission. Another option is to adopt the strategy of some courses administering a "dummy TMA", where students try the process of submitting an assignment. This might, however, cause frustration for experienced students if it is mandatory. Lastly, a questionnaire collecting student preferences might identify attitudes to study online, downloaded or printed materials. Knowing this data might decrease some students being flagged as at-risk and avoid unnecessary contact by a tutor when it is not needed.

5.2. Limitations

Limitations of this study could be addressed in future research. Our student selection meant that we had to narrow it down to first-year students and the first assignment in the course. This is where the models' errors are expected to be higher, as there is not as much collected information about the progress of students to inform predictions. The errors in the later phases of their studies might have different characteristics and require further examination, so future research can deal with that. Moreover, it is possible that due to the small self-selected sample, our study may have attracted students who tended to explain their study progress mainly based on external factors, i.e. family responsibilities or computer issues, rather than internal factors, i.e. being reflective of how their own study patterns, study dedication, motivation or procrastination (Imhof et al., 2021). Another limitation is that we only collected data from one faculty (STEM), whereas examining other courses might enrich our understanding of why errors in predictions occur.

6. Conclusions

This study has focused on the qualitative analysis of machine learning errors of PLA models of OUAnalyse - a predictive system used at The Open University - for the first assignment of Level 1 STEM courses. In particular, we looked at why predictions with high confidence turn up the other way. Acknowledging that existing quantitative data cannot explain such patterns, we conducted in-depth semi-structured interviews with 27 students, both predictions of at-risk students submitting their assignments (false positives) and the ones predicted as on track but not submitting (false negatives).

Findings revealed the significance of unexpected events occurring during studies that can affect students' behaviour and cannot be foreseen and accounted for in PLA, such as changes in family and work responsibilities, unexpected health issues and computer problems. Interview data helped identify new data sources, which could be integrated into predictions to mitigate some of the errors, such as study loan application information. Other changes in the design, e.g. making tutorials available online or changing assessment weighting strategy, can remove existing barriers encouraging student engagement and making the VLE more inclusive. This is in line with studies noting the importance of learning design, explaining 69% of the variance in student behaviour (Rienties, Nguyen, Holmes, & Reedy, 2017).

Together with our previous quantitative study (Hlosta, Papathoma, & Herodotou, 2020), the results show the benefits of the integrative approach of mixed-method research, when the qualitative research complements the findings from the large-scale quantitative analysis. Moreover, it shows the importance of including students in the research not only in the design of the predictive technology but collecting their feedback after deployment in the production.

Further, insights from this study showcase the importance of complimenting AI-based systems with human intelligence, in our case, teachers, that is over and above what a good AI system can offer. Teachers may be aware of unexpected events a student is facing as they have direct communication with their students, and they can thus provide guidance, encourage tutorial participation and suggest seeking help from student support services if needed. Teachers should be aware that errors may occur in models and use model data in conjunction with other information they may have from their direct communication with students to support students as best as they can.

Finally, these insights raise the need for teacher training to highlight good practices of the right timing and the ways of getting in touch with students, which should ideally be part of a university-wide policy on the use of PLA. Such contact should support developing an academic connection and social presence, which are previously shown to facilitate persistence in studying (Lukosius, Pennington, & Olorunniwo, 2013; Milem & Berger, 1997). The fact that even high accuracy PLA manifests some errors should not discourage its use. An increasing body of evidence points to PLA's role in increasing student pass and completion rates as well as teaching practice when used systematically by teachers (Herodotou et al., 2019; Herodotou, Maguire, McDowell, & Hlosta, 2021), while they can promote equity in education by benefiting in particular economically disadvantaged students (Hlosta, Herodotou, Bayer, & Fernandez, 2021). As noted by Kitto et al. (2018, p. 456): "Overemphasising computational accuracy is likely to delay the adoption of LA tools that could already be used productively."

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the internal esteem grant [Disproved predictions of at-risk students: Some students fail despite doing well, others succeed despite predicted as at-risk]; by The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.caeai.2022.100108.

References

Alexandron, G., Yoo, L. Y., Ruipérez-Valiente, J. A., Lee, S., & Pritchard, D. E. (2019). Are MOOC learning analytics results trustworthy? With fake learners, they might not Be. *International Journal of Artificial Intelligence in Education, 29*(4), 484–506. https://doi.org/10.1007/s40593-019-00183-1

Anderson, H., Boodhwani, A., & Baker, R. (2019). *Assessing the fairness of graduation predictions*. EDM.

Archer, E., & Prinsloo, P. (2020). Speaking the unspoken in learning analytics: Troubling the defaults. *Assessment & Evaluation in Higher Education, 45*(6), 888–900. https://doi.org/10.1080/02602938.2019.1694863

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 267–270). https://doi.org/10.1145/2330601.2330666

Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00285-9

Baker, R., Xu, D., Park, J., Yu, R., Li, Q., Cung, B., … Smyth, P. (2020). The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: Opening the black box of learning processes. *International Journal of Educational Technology in Higher Education, 17*(1), 1–24.

Bannert, M., Molenaar, I., Azevedo, R., Järvelä, S., & Gašević, D. (2017). Relevance of learning analytics to measure and support students' learning in adaptive educational technologies. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 568–569. https://doi.org/10.1145/3027385.3029463

Bayer, V., Hlosta, M., & Fernandez, M. (2021). Learning analytics and fairness: Do existing algorithms serve everyone equally? In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Artificial intelligence in education* (pp. 71–75). Springer International Publishing. https://doi.org/10.1007/978-3-030-78270-2_12.

Beutel, M. E., Klein, E. M., Aufenanger, S., Brähler, E., Dreier, M., Müller, K. W., et al. (2016). Procrastination, distress and life satisfaction across the age range – a German representative community study. *PLoS One, 11*(2), Article e0148054. https://doi.org/10.1371/journal.pone.0148054

Blanzeisky, W., & Cunningham, P. (2021). *Algorithmic factors influencing bias in machine learning*. ArXiv:2104.14014 [Cs, Stat] http://arxiv.org/abs/2104.14014.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Britt, S. L., Ammerman, D. A., Barrett, S. F., & Jones, S. (2017). Student loans, financial stress, and college student retention. *The Journal of student financial aid, 47*(1), 3.

Calvert, C., & Hilliam, R. (2019). Student feedback to improved retention: Using a mixed-methods approach to extend specific feedback to a generalisable concept. *Open Learning: The Journal of Open, Distance and e-Learning, 34*(1), 103–117. https://doi.org/10.1080/02680513.2018.1552580

Chmielewski, A. K. (2019). The global increase in the socioeconomic achievement gap, 1964 to 2015. *American Sociological Review, 84*(3), 517–544.

De Gayardon, A., Callender, C., & Green, F. (2019). The determinants of student loan take-up in England. *Higher Education, 78*(6), 965–983.

De Laet, T., Millecamp, M., Broos, T., De Croon, R., & Verbert, K. (2020). Explainable learning analytics: Challenges and opportunities. In *Companion proceedings of the 10th international conference on learning analytics & knowledge LAK20* (pp. 500–510). Society for Learning Analytics Research ((SoLAR)).

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice, 13*(1), 17–35.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Ding, M., Wang, Y., Hemberg, E., & O'Reilly, U.-M. (2019). Transfer learning using representation learning in massive open online courses. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 145–154. https://doi.org/10.1145/3303772.3303794

Ding, M., Yang, K., Yeung, D.-Y., & Pong, T.-C. (2019). Effective feature learning with unsupervised learning for improving the predictive models in massive open online courses. In *Proceedings of the 9th international conference on learning analytics & knowledge*.

Fei, M., & Yeung, D.-Y. (2015). Temporal models for predicting student dropout in massive open online courses. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. https://doi.org/10.1109/ICDMW.2015.174, 256–263.

García-Martín, E., & Lavesson, N. (2017). Is it ethical to avoid error analysis?. ArXiv: 1706.10237 [Cs] http://arxiv.org/abs/1706.10237.

Gardner, J., Yang, Y., Baker, R., & Brooks, C. A. (2019). *Modeling and experimental design for MOOC Dropout prediction: A replication perspective*. EDM.

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education, 28*, 68–84.

Greenwell, B., Boehmke, B., & Cunningham. (2019). Package 'gbm. *R package version, 2* (5).

Herodotou, C., Hlosta, M., Boroowa, A., Rienties, B., Zdrahal, Z., & Mangafa, C. (2019). Empowering online teachers through predictive learning analytics. *British Journal of Educational Technology, 50*(6). https://doi.org/10.1111/bjet.12853

Herodotou, C., Maguire, C., McDowell, N., & Hlosta, M. (2021). The engagement of university teachers with predictive learning analytics. Computers & Education,. *Computers & Education, 173*. https://doi.org/10.1016/j.compedu.2021.104285

Hlosta, M., Herodotou, C., Bayer, V., & Fernandez, M. (2021). Impact of Predictive Learning Analytics on Course Awarding Gap of Disadvantaged Students in STEM. *In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), Artificial Intelligence in Education (pp. 190–195). Springer International Publishing.*. https://doi.org/10.1007/978-3-030-78270-2_34

Hlosta, M., Zdrahal, Z., & Zendulka, J. (2017). Ouroboros: Early identification of at-risk students without models based on legacy data. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 6–15. https://doi.org/10.1145/3027385.3027449

Hlosta, M., Papathoma, T., & Herodotou, C. (2020). Explaining Errors in Predictions of At-Risk Students in Distance Learning Education. *I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), Artificial Intelligence in Education*, 119–123.

Hlosta, M., Zdrahal, Z., Bayer, V., & Herodotou, C. (2020). Why Predictions of At-Risk Students Are Not 100% Accurate? Showing Patterns in False Positive and False Negative Predictions. *Companion Proceedings of the 10th International Learning Analytics & Knowledge Conference. Frankfurt, Germany 2020*.

Imhof, C., Bergamin, P., & McGarrity, S. (2021). Prediction of dilatory behaviour in online assignments. *Learning and Individual Differences, 88*, Article 102014. https://doi.org/10.1016/j.lindif.2021.102014

Kitto, K., Shum, S. B., & Gibson, A. (2018). Embracing imperfection in learning analytics. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 451–460. https://doi.org/10.1145/3170358.3170413

Kizilcec, R. F., & Lee, H. (2021). *Algorithmic fairness in education. ArXiv:2007.05443 [Cs]* http://arxiv.org/abs/2007.05443.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), *8th innovations in theoretical computer science conference (ITCS 2017): Vol. 67. Schloss Dagstuhl–Leibniz-Zentrum fuer informatik*. https://doi.org/10.4230/LIPIcs.ITCS.2017.43, 43:1-43:23).

Klingler, S., Wampfler, R., Käser, T., Solenthaler, B., & Gross, M. H. (2017). Efficient feature embeddings for student classification with variational auto-encoders. *Proceedings of the 10th International Conference on Educational Data Mining*, 72–79.

Knowles, J. E. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining, 7*(3), 18–67. https://doi.org/10.5281/zenodo.3554725

Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrahal, Z., Vaclavek, J., & Wolff, A. (2015). OU analyse: Analysing at-risk students at the open university. *Learning Analytics Review, LAK15–1*, 1–16.

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., et al. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1909–1918). https://doi.org/10.1145/2783258.2788620. Association for Computing Machinery.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM conference on AI* (pp. 131–138). Ethics, and Society.

Lee, H., & Kizilcec, R. F. (2020). *Evaluation of fairness trade-offs in predicting student success*. ArXiv:2007.00088 [Cs] http://arxiv.org/abs/2007.00088.

Lukosius, V., Pennington, J. B., & Olorunniwo, F. O. (2013). How students' perceptions of support systems affect their intentions to drop out or transfer out of college. *Review of Higher Education and Self-Learning, 6*(18), 209–221.

Mathrani, A., Susnjak, T., Ramaswami, G., & Barczak, A. (2021). Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. *Computers and Education Open, 2*, Article 100060. https://doi.org/10.1016/j.caeo.2021.100060

Milem, J. F., & Berger, J. B. (1997). A modified model of college student persistence: Exploring the relationship between Astin's theory of involvement and Tinto's theory of student departure. *Journal of College Student Development, 38*(4), 387–400.

Ochoa, X., & Merceron, A. (2018). Quantitative and qualitative analysis of the learning analytics and knowledge conference 2018. *Journal of Learning Analytics, 5*(3), 154–166. https://doi.org/10.18608/jla.2018.53.10

Prenkaj, B., Distante, D., Faralli, S., & Velardi, P. (2021). Hidden space deep sequential risk prediction on student trajectories. *Future Generation Computer Systems, 125*, 532–543.

Qiu, J., Tang, J., Liu, T. X., Gong, J., Zhang, C., Zhang, Q., et al. (2016). Modeling and predicting learning behavior in MOOCs. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 93–102. https://doi.org/10.1145/2835776.2835842

Rets, I., Herodotou, C., Bayer, V., Hlosta, M., & Rienties, B. (2021). Exploring critical factors of the perceived usefulness of a learning analytics dashboard for distance university students. *International Journal of Educational Technology in Higher Education, 18*(1), 1–23.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should i trust you?' Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rienties, B., Nguyen, Q., Holmes, W., & Reedy, K. (2017). A review of ten years of implementation and research in aligning learning design with learning analytics at the Open University UK. *Interaction Design and Architecture (s), 33*, 134–154.

Riestra-González, M., Paule-Ruíz, M. del P., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers & Education, 163*, Article 104108. https://doi.org/10.1016/j.compedu.2020.104108

Riveiro, M., & Thill, S. (2021). That's (not) the output I expected!" on the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence, 298*, Article 103507. https://doi.org/10.1016/j.artint.2021.103507

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review, 3* (1), 16–20.

Schumacher, C., & Ifenthaler, D. (2018). Features students really expect from learning analytics. *Computers in Human Behavior, 78*, 397–407. https://doi.org/10.1016/j.chb.2017.06.030

Springer, A., & Whittaker, S. (2018). *'I had a solid theory before but it's falling apart': Polarizing Effects of Algorithmic Transparency*, Article 02163. *ArXiv, abs/1811*.

Tang, S., Peterson, J. C., & Pardos, Z. A. (2016). Deep neural Networks and how they apply to sequential education data. In *Proceedings of the Third (2016) ACM conference on learning @ scale* (pp. 321–324). https://doi.org/10.1145/2876034.2893444

Taylor, C., Veeramachaneni, K., & O'Reilly, U.-M. (2014). *Likely to stop? Predicting stopout in massive open online courses. ArXiv:1408.3382 [Cs]* http://arxiv.org/abs/1408.3382.

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior, 47*, 157–167.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89–125. https://doi.org/10.3102/00346543045001089

Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Van Assche, F., Parra, G., et al. (2014). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing, 18*(6), 1499–1514. https://doi.org/10.1007/s00779-013-0751-2

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law, 7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR (SSRN scholarly paper ID 3063289). *Social Science Research Network*. https://doi.org/10.2139/ssrn.3063289

Walker-Gibbs, B., Ajjawi, R., Rowe, E., Skourdoumbis, A., Thomas, M. K. E., O'Shea, S., et al. (2019). *Success and failure in higher education on uneven playing fields*. Curtin University.

Wigfield, A., & Eccles, J. S. (2002). *The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence*. Development of achievement motivation.

Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes? *Proceedings of the Eighth ACM Conference on Learning @ Scale*, 91–100. https://doi.org/10.1145/3430895.3460139

Yu, R., Li, Q., Fischer, C., Doroudi, S., & Xu, D. (2020). Towards accurate and fair prediction of college success: Evaluating different sources of student data. In *International educational data mining society*. International Educational Data Mining Society.

First Author, Second Author, 2020a.

First Author, Third Author, Second Author, 2020b.