iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IX **Month of publication:** September 2026

DOI: https://doi.org/10.22214/ijraset.2026.84783

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IX Sep 2026- Available at www.ijraset.com

Reimagining Education: Innovations in Teaching, Learning and AI in Higher Education

Perseus Bhavnagri[1], Shraddha Ayare Shirke[2]
[1]*Department of Information Technology, Wilson College, University of Mumbai*
[2]*Faculty, Master of Science Information Technology, Wilson College*

Abstract: ***Universities are being offered artificial intelligence that flags students likely to fail, and the case for buying it rests on accuracy figures drawn from papers that split one pooled pile of student records at random. This study asks how much of that accuracy a university would actually see. Using the Open University Learning Analytics Dataset, 32,593 student registrations across seven modules and twenty-two presentations with 10,655,280 recorded interactions, a gradient-boosted model was trained to predict failure or withdrawal from what was known on day 30 of a course. Under a random split it reaches an area under the ROC curve of 0.8450. Three sources of inflation were then separated on identical test rows. Losing familiarity with the course costs almost nothing, 0.0085 of area under the curve, consistent in direction across 17 of 22 presentations but small by every effect size used (Wilcoxon p = 0.00219, Cliff's delta 0.223, bootstrap interval spanning zero); training on 2013 and predicting 2014 costs 0.0037 and is not significant (p = 0.12720). What does matter is arithmetic rather than transfer. Pooling students across presentations that fail at rates from 0.274 to 0.658 adds 0.0267 that no single course ever sees, and a model told only which presentation a student is enrolled on, nothing about the student, already reaches 0.5777. Larger still, 15.7 percent of students had already unregistered before day 30; removing them costs 0.0575. The model is better at recognising students who have already left than students who will fail while trying (0.8799 against 0.7948). At an alert budget of one student in five it misses 58.6 percent of those at risk, and misses the least deprived more often than the most deprived, a pooled gap of 0.118 that shrinks to 0.086 and reverses in one module once deprivation is compared within courses. Applied together, an unseen course and a population restricted to students still enrolled bring the figure to 0.7520, a fall of 0.0931 of which 62.2 percent is departed students, 28.7 percent is pooling and 9.1 percent is the unfamiliar course. The paper concludes that reported accuracy in this literature is largely a property of how the test set was assembled, and recommends within-course reporting, the exclusion of students already gone, and a stated alert budget.***

Keywords: ***learning analytics, early-warning systems, student retention, at-risk prediction, artificial intelligence in higher education, evaluation methodology, model generalisation, algorithmic fairness, Simpson's paradox, educational data mining.***

I. INTRODUCTION

The phrase reimagining education now usually arrives attached to a product. Among the products, the one universities have bought earliest and most widely is not a chatbot or a marking assistant but a predictor: software that watches what students do in a virtual learning environment and tells staff which of them are heading for failure, early enough that somebody can telephone. The idea is old enough to have a track record and new enough to still be sold as artificial intelligence, and it is attractive for a reason that has nothing to do with novelty. Contacting a struggling student in week four is cheap. Losing them in week thirty is not.

The evidence offered for these systems is a number, and the numbers are encouraging. On the dataset used in this paper, a recent study reports a model that catches 97 percent of the students who will fail, as early as week two of a course [1]. On a school system covering 451,852 students, another reports one that correctly flags 86.7 percent of those who will fail, five months before the year ends [2]. Read quickly, a figure like that is a promise about what will happen when the software is switched on. Read carefully, it is a statement about a particular set of test rows, and the two are not the same thing. The gap between them is not a matter of anyone cheating. It is that the number answers a different question from the one the university is asking.

An accuracy figure produced by a random split answers this: if I take all the student records I have, from every course and every year, shuffle them, hide a fifth of them, and train on the rest, how well can I rank the hidden fifth? A university deploying the same model asks something else: I have last year's records; how well will this rank the students sitting in this course, right now, many of whom are enrolled on a module the model has never seen? The two questions have the same words in them and different answers, and the gap between the answers is the subject of this paper.

The gap could come from several places, and it matters which. It could be that student behaviour means different things in different courses, so a model tuned on six modules misreads the seventh. It could be that cohorts drift, so last year's model is stale. It could be that pooling many courses into one test set flatters the metric by adding variation between courses that no single course contains. Or it could be that a large part of what the model appears to detect is not a student in difficulty at all, but a student who has already gone, whose file closed weeks before the prediction was made and who cannot be helped by any alert. These four explanations imply four different fixes, and only one of them is a modelling problem.

This study separates them. Working with the Open University Learning Analytics Dataset, a public release covering 32,593 student registrations on seven distance learning modules with 10,655,280 recorded interactions, it builds one at-risk model in the ordinary way and then evaluates it four different ways on identical test rows, so that the only thing changing is the design of the evaluation. The result is a decomposition: how much of the headline figure survives each move towards realism, and what a university should expect instead.

The answer is not the one the study set out expecting. The portability of these models turns out to be good. Handing a model a course it has never seen barely hurts it, and asking last year's model to rank this year's students hurts it less still. What inflates the reported figure is quieter and harder to argue with: the way the test set is put together, and the presence in it of students the system exists to reach and has already lost.

II. REVIEW OF LITERATURE

A. The dataset this literature is built on

Kuzilek, Hlosta and Zdrahal released the Open University Learning Analytics Dataset in 2017 to give the field a shared, citable resource with both demographic and behavioural records in it [3]. The Open University teaches at a distance, so almost everything a student does with course material passes through the virtual learning environment and is logged; the release pairs those logs with declared background, assessment marks and a final outcome. Seven modules were chosen from a much larger catalogue on the criteria that each had at least 500 students, ran more than once, was taught through the virtual learning environment, and failed a meaningful number of people. The authors anonymised the file with the ARX tool to a k-anonymity of five, which is what reduced the population from 38,239 students to the 32,593 published, and which is why age and deprivation arrive as bands rather than values.

The release has since become a benchmark, to the point that Howard published an R package whose stated purpose is to spare researchers the repeated work of reshaping it, and whose existence is itself evidence of how many teams reshape it [4]. Work by the Open University's own group used the same underlying records to model student activity with unary hypothesis automata and Markov chains, arguing that the sequence of what a student touches carries information that a total does not [5].

B. What has been reported on it

Prediction results on this dataset are plentiful and generally good. Kaushal and Mall compared decision trees against a long short-term memory network at six points in a course and reported recall of 0.97 as early as week two, concluding that prediction is possible before any assessment has been marked [1]. Their more striking claim is about what carries the signal: they attribute 68 percent of feature importance to static demographic and pre-enrolment attributes.

Da Silva, Eicher and Longo built a survival-analysis benchmark on the same data and reached the opposite conclusion on that specific point, finding across model families that the dominant signal was temporal and behavioural rather than background [6]. Their paper is also unusually careful about comparison: it declines to rank models across its two families on the grounds that the ranking would confuse genuine differences with artefacts of how time was represented. That caution is close to the concern of the present study, arrived at from a different direction. Labba and Boyer, working on the same dataset, addressed the fact that educational data arrives as a stream rather than a block, and reported roughly a ten percent gain in accuracy from a genetic-algorithm approach to choosing which samples to retrain on [7].

It is worth being specific about how these results are produced, since that is what this paper examines. Da Silva and colleagues partition at the level of the enrolment, stratified by whether the student dropped out, and their own audit confirms that the resulting split spans all seven modules and all twenty-two module-presentations [6]. Kaushal and Mall go further in the permissive direction, training a single model on the full course record and then testing it on earlier snapshots of the same students, a design they adopt deliberately so that predictions at a given week use only the information available by that week [1]. Neither protocol is careless, and neither is unusual. Both put students from every course into one pile before the split, which is exactly the arrangement whose value is measured below.

Outside this dataset the picture is similar. Aulck, Velagapudi, Blumenstock and West predicted dropout from the transcript and demographic records of more than 32,500 students enrolled at a very large American state institution, and Gardner and Brooks surveyed the whole business of success prediction in massive open online courses, giving particular attention to how prediction problems are framed and evaluated [8], [9].

C. Whether these models travel

The most direct precedent for this paper is Gasevic, Dawson, Rogers and Gasevic, who examined nine blended undergraduate courses and asked whether one predictive model could serve them all [10]. It could not, in an interesting way: which log variables predicted success, and how strongly, changed from course to course according to how the teaching used the technology. Their conclusion was that a generalised model can misstate the effect of a feature in either direction, and that instructional context has to be part of the analysis rather than noise around it. The present study takes that finding as its starting point and asks a narrower, quantitative question the original did not: not whether the coefficients change, but how much discrimination is actually lost when the model meets a course it has never seen.

Swamy, Marras and Kaser attacked the same problem as an engineering task, developing transfer strategies so that a model trained on a broad set of courses can be started up on a new one, and reporting on 26 courses with more than 145,000 enrolments that a model combining interaction data with course-level information can match one that has seen earlier runs of the same course [11]. Their result and the one below point the same way. Theirs shows that transfer can be made to work; this one measures how much is at stake if nothing special is done.

Medicine has had this argument already and settled it more firmly. Futoma, Simons, Panch, Doshi-Velez and Celi argued that generalisability in clinical machine learning is often assumed rather than demonstrated, and that a model validated where it was born tells you little about a model deployed elsewhere [12]. Roberts and colleagues reviewed hundreds of models built to diagnose or prognosticate COVID-19 from chest images and found almost none fit for clinical use, with evaluation design among the recurring faults [13]. Education has not yet produced an equivalent audit, and the modest cost of transfer found here suggests the education case may be genuinely less severe than the clinical one, rather than merely less examined.

D. Fairness, and what a threshold does

Kizilcec and Lee set out how the fairness literature maps onto education, including the awkward fact that the standard criteria cannot all be satisfied at once when base rates differ between groups [14]. That impossibility is not abstract here: the groups compared below do differ in base rate, and the pattern of misses reported in Section VI is a direct consequence.

Yu, Lee and Kizilcec asked whether dropout models should be given protected attributes at all, and found the answer is not the obvious one, since removing a variable does not remove the information [15]. Riazy, Simbeck and Schreck audited at-risk prediction in a virtual learning environment for discrimination and proposed checks to run on the data before modelling begins [16]. Anahideh, Haghighat, Nezami and Gandara showed that a decision made long before the classifier, namely how missing values are filled, changes the measured fairness of the result [17]. Bird, Castleman and Song examined racial bias in community college success prediction and made the point that matters for deployment: bias in a model becomes injustice at the moment a threshold turns a score into an action, because that is when scarce support is handed out [18].

Jayaprakash, Moody, Lauria, Regan and Baron reported one of the few open accounts of an early alert system in actual use, including the part most papers omit, which is what staff did with the alerts [19]. Their study also asked the portability question directly, moving a model built at a private liberal arts college to partner institutions including two-year community colleges, which makes it the earliest of the works cited here to treat transfer as something to be measured rather than assumed. Baker, Bosch, Hutt, Zambrano and Bowers defended the area under the ROC curve against a broad critique, arguing that most complaints about the metric are really complaints about using any single metric alone [20]. Their argument is accepted here without reservation; this paper does not claim the metric is broken, only that its value depends on which rows are in the test set. El Jihaoui, Abra and Mansouri built an early warning system for 451,852 school students under an explicit recall constraint and audited it across intersecting groups, an approach that treats the alert budget as part of the problem rather than an afterthought [2].

Zawacki-Richter, Marin, Bond and Gouverneur reviewed the wider field of artificial intelligence in higher education and found remarkably little participation from educators and remarkably little discussion of ethics in the technical literature [21]. Their observation frames this paper's motivation: decisions about which students get help are being settled inside evaluation protocols that the people affected never see.

E. Statistical practice

Lakens set out why an effect size must be reported next to a p value, since with tens of thousands of rows significance is nearly guaranteed and says nothing about whether a difference is worth acting on [22]. That warning is load-bearing here: the central comparison in Section VI is significant and small at the same time, and the paper's conclusion depends on saying so. Kievit, Frankenhuis, Waldorp and Borsboom gave a practical account of Simpson's paradox, in which a relationship measured across a pooled sample differs from the relationship inside every subgroup [23]. Their guidance is applied in Section VI-G as a routine check rather than as a reaction to a surprising number.

III. GAP ANALYSIS

Three gaps sit between what has been published and what a university needs to know.

1) The portability question has been answered qualitatively but not quantitatively. Gasevic and colleagues established that predictors differ between courses [10], and Swamy and colleagues built methods to transfer models between them [11]. Neither reports what it costs, in the metric everyone quotes, to hand an ordinary model a course it has never seen. Without that number a university cannot discount a published figure.

2) The population being predicted is rarely stated. Prediction studies on this dataset generally treat failure and withdrawal as one outcome and keep every registration in the sample. That includes students who had already unregistered before the prediction date. No published study on this data appears to report what the accuracy figure becomes once those students are removed, although they are precisely the students an alert cannot reach.

3) The accuracy figure is reported pooled, and the pooling is not examined. Module-presentations in this dataset fail between 27.4 and 65.8 percent of their students, and the published splits are taken across all of them at once [6]. A metric computed that way contains variation that no single course contains, and no study located here separates that contribution from the model's ability to rank students within a course.

A fourth, smaller gap concerns a live disagreement. Kaushal and Mall report that static background attributes carry roughly two thirds of the predictive weight [1]; da Silva and colleagues report on the same dataset that behaviour dominates [6]. The measurement in Section VI-H speaks to that disagreement directly.

IV. OBJECTIVES, VARIABLES AND HYPOTHESES

A. Objectives

1) To measure what an at-risk model reported under a random split is worth when the same model meets a course it has never seen, and when it meets a cohort a year later than the one it learned from.

2) To separate the contribution of pooling students across courses from the model's ability to rank students within a course.

3) To measure how much of the reported figure comes from students who had already left before the prediction was made.

4) To report what an alert actually achieves at a support budget a university could staff, and whether the students it misses are distributed evenly.

5) To publish a deployment-realistic benchmark figure that later work can be compared against.

B. Variables

The outcome is binary and follows the convention of this literature: a student is at risk if the recorded final result is Fail or Withdrawn, and not at risk if it is Pass or Distinction. Across the whole file 52.80 percent of registrations are at risk.

The predictors are the 39 columns a university would hold on day 30 of a presentation, in three groups. Declared attributes, recorded at registration: gender, region, highest previous education, deprivation band, age band, number of previous attempts at the module, credits being studied, declared disability and registration date. Behaviour in the virtual learning environment up to and including day 30: total clicks, click events, active days, distinct materials touched, largest single day, clicks before the module started, clicks in each of the first four weeks, clicks by material type for the eight commonest types, days since the last click, clicks per active day, and whether the student engaged before the start date at all. Assessment: how many pieces of work were submitted by day 30, their mean and minimum marks, how many were carried over from a previous presentation, whether any mark was below the 40 point pass threshold, whether nothing was submitted, and how many assessments the module had scheduled by that date.

The evaluation design is the variable of interest. It takes four values: a random split across all rows; a held-out presentation with the model having seen other presentations of the same module; a held-out presentation with the module withheld entirely; and training on 2013 with testing on 2014.

C. *Hypotheses*

These were written down after the twenty-three sources were read and after the dataset was inspected, but before any model was fitted. The record is in Hypotheses (pre-registered).md alongside the analysis scripts.

1) H1. The area under the curve from a random split is materially higher than the area under the curve when the model is tested on a module it has never seen.
2) H2. The area under the curve from a random split is materially higher than the area under the curve when the model is trained on 2013 and tested on 2014.
3) H3. A large share of the pooled figure comes from separating students who have already stopped rather than students who are still studying. Tested in two parts: whether module identity alone is informative, and whether restricting the population lowers the figure.
4) H4. At one global cut-off, the proportion of genuinely at-risk students who are missed differs across deprivation bands and levels of prior education.
5) H5. Any disparity found under H4 holds in the same direction inside each module separately.

The pre-registration also fixed, in advance, that a small or absent effect would be reported as such. Two of the five hypotheses below are not supported, and the paper's contribution rests partly on saying so.

V. RESEARCH METHODOLOGY

A. *Research design*

This is a secondary quantitative study of an existing public dataset, conducted in September 2026. No new data were collected and no human participants were approached. The design is a within-subjects comparison of evaluation protocols: the same students, the same features and the same learning algorithm are held fixed while the rule for choosing training data is varied, so that any difference in the measured metric is attributable to the evaluation design and to nothing else.

B. *Data source*

The Open University Learning Analytics Dataset was downloaded on 1 September 2026 from the UCI Machine Learning Repository, entry 349, under a Creative Commons Attribution 4.0 licence. The official Open University download address returns a web page rather than the archive when requested by a script, so the UCI mirror of the same certified release was used. The archive holds 32,593 registrations, 10,655,280 click records, 173,912 assessment submissions, 206 assessment definitions and 6,364 catalogued materials, across seven modules and twenty-two module-presentations from 2013 and 2014. Presentations beginning in February carry the suffix B and those beginning in October carry J. Two properties of the file shape the analysis and are recorded rather than smoothed over. Module CCC ran only in 2014 and module AAA only in October, so a comparison that trains on 2013 and tests on 2014 mixes the effect of a new year with the effect of a new course; that comparison is therefore reported both including and excluding CCC. And one deprivation band is written in the source file without its per cent sign, which is a typing slip rather than a separate category and is joined to the correctly written band.

C. *Prediction task*

The prediction is made on day 30 of a presentation, counted from the module start date. Day 30 is late enough for a month of behaviour to exist and early enough for an intervention to be worth making. Everything recorded after day 30 is excluded from the feature table, including from the construction of categories, so no information from the future can leak backwards. Of the 10,655,280 click records, 2,980,575 fall on or before day 30, along with 26,522 assessment submissions.

A student with no click record is treated as having clicked zero times rather than as missing, because on this dataset the absence of a record is itself the measurement. By day 30, 3,751 students had not clicked at all and 11,289 had submitted nothing.

D. *Models*

Two classifiers are used: penalised logistic regression, and histogram-based gradient-boosted trees with 300 iterations at a learning rate of 0.06. Categorical columns are one-hot encoded for the linear model and passed as native categories to the boosted model. Numeric columns are median-imputed and standardised for the linear model; the boosted model reads missing values directly, which matters because a missing assessment mark is meaningful here.

Neither model is tuned separately for each evaluation design. Tuning per design would make the designs incomparable, which would defeat the purpose. The boosted model is used for every comparison after Section VI-A because it is the stronger of the two and therefore the more favourable to the systems under examination.

E. Evaluation designs

1) Random split. Stratified five-fold cross-validation over all 32,593 rows, with out-of-fold predictions pooled into a single figure. This reproduces the protocol behind most published numbers.

2) Course seen. For each of the twenty-two presentations in turn, the model is trained on every other row in the dataset, including other presentations of the same module, and tested on that presentation.

3) Course unseen. For the same twenty-two presentations and exactly the same test rows, the model is trained only on rows belonging to other modules, so the module itself is entirely withheld.

4) Forward in time. The model is trained on the 2013 presentations and tested on the 2014 ones, and separately compared against a contemporaneous model on identical test rows.

The second and third designs share their test rows exactly, which makes the comparison paired and removes any possibility that the difference reflects which students happened to fall in which fold.

F. Statistical methods

Discrimination is reported as the area under the ROC curve, computed through the rank identity so that ties contribute one half, alongside average precision and the recall achieved when the highest-scoring 20 percent of students are flagged. That last figure is the operational one: a support team has capacity for a fixed number of conversations, not for everyone the model dislikes.

Paired comparisons across presentations use the Wilcoxon signed-rank test with Cliff's delta as the effect size, both distribution-free, and a percentile bootstrap on the difference of means with 10,000 resamples. Confidence intervals on a single area under the curve are computed twice, once resampling rows and once resampling whole presentations, because students inside one presentation are not independent observations and the row-level interval is too narrow. Differences in proportions use a two-proportion z test with the risk difference, its interval and an odds ratio; contingency tables use the chi-square test with Cramer's V. The random seed is 20260901, set once, and every figure in this paper reproduces exactly on a rerun.

VI. DATA ANALYSIS AND FINDINGS

A. How much the courses differ before any model is fitted

The seven modules are not variations on one course. They fail between 29.0 and 62.2 percent of their students, and across the twenty-two presentations the range widens to 27.4 to 65.8 percent, a spread of 38.3 percentage points. The share of students who had unregistered before day 30 varies almost as much, from 2.9 percent on GGG-2013J to 23.1 percent on CCC-2014B. Table I gives the modules and Fig. 1 the presentations.

TABLE I
THE SEVEN MODULES

Module	Domain	Presentations	Students	At-risk rate
AAA	Social Sciences	2	748	0.290
BBB	Social Sciences	4	7,909	0.525
CCC	STEM	2	4,434	0.622
DDD	STEM	4	6,272	0.584
EEE	STEM	3	2,934	0.438
FFF	STEM	4	7,762	0.530
GGG	Social Sciences	3	2,534	0.403

Domains are as published with the dataset. At-risk means a final result of Fail or Withdrawn. CCC ran only in 2014.

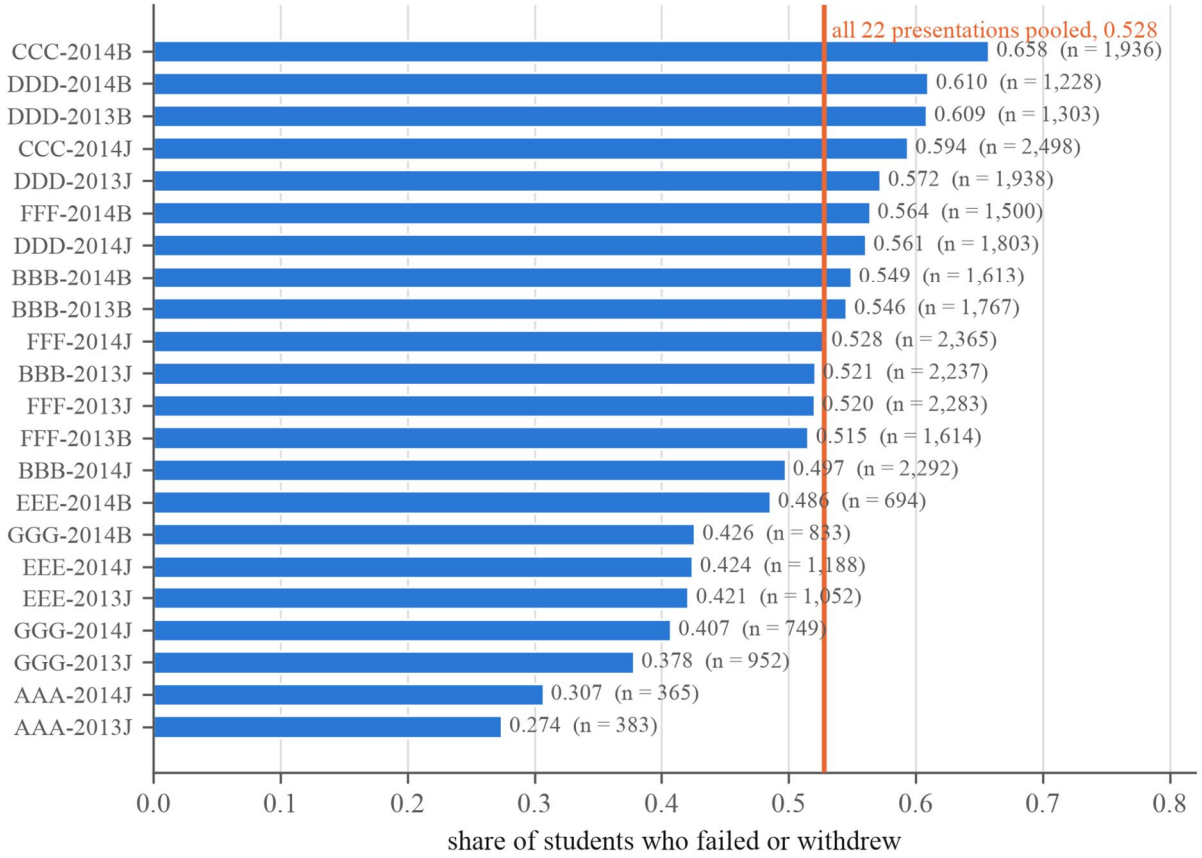


Fig. 1 How much the modules differ before any model is fitted.

B. The figure a random split produces

Trained and tested in the ordinary way, the gradient-boosted model reaches an area under the ROC curve of 0.8450 with average precision 0.8799.

The logistic model reaches 0.8231 and 0.8599. Resampling rows gives a 95 percent interval of 0.8405 to 0.8491 for the boosted model; resampling whole presentations, which is the honest unit here, widens it to 0.8291 to 0.8596. Adding the presentation identity as an explicit feature moves the figure only to 0.8484, which is the first hint that the model is already recovering course context from behaviour without being told it.

These are day-30 figures from features that stop at day 30, and they are not offered as a competitive result. The nearest published comparison, a long short-term memory model on this same dataset, reports an overall area under the curve of 0.7678 [1], which is lower; other work on the dataset reports recall, accuracy or concordance rather than this metric [6], [7], so a like-for-like ranking is not available and is not attempted. Nothing below compares this study's figure to anyone else's. Every comparison in this paper is internal, against itself, with the evaluation design as the only thing that moves.

C. What a model knows from the course alone

The first part of H3 asks how much of the pooled figure is a statement about courses rather than about students. A model was fitted whose only input is which module-presentation a student is enrolled on. It receives no demographics, no clicks and no marks; it can do nothing but assign every student on a presentation the historical failure rate of that presentation.

That model reaches an area under the curve of 0.5777. Measured above chance, it covers 22.5 percent of the distance the full model covers. Close to a quarter of the apparent discrimination in the headline figure is therefore available without knowing anything whatsoever about the individual student, and is a consequence of pooling students from courses that fail at different rates. H3a is supported.

D. The same students, a model that has and has not seen their course

H1 is the central test. For each of the twenty-two presentations the test rows are held identical and only the training set changes: in one condition the model has seen other presentations of the same module, in the other the module is withheld entirely.

Averaged across presentations, the model that has seen the course reaches 0.8183 and the model that has not reaches 0.8099. The mean cost of an unfamiliar course is 0.0085 of area under the curve. The direction is consistent, with 17 of the 22 presentations worse when the module is withheld, and the Wilcoxon signed-rank test on the pair is significant at W = 36.0, p = 0.00219. The magnitude is not. Cliff's delta between the two sets of areas is 0.223, which is small, and a bootstrap on the difference of means gives 0.0085 with an interval from -0.0276 to 0.0448, spanning zero. At the operational end the difference disappears entirely: recall at a 20 percent alert budget is 0.370 with the course seen and 0.369 without it. Fig. 2 shows every presentation, paired.

H1 as written is therefore not supported. The direction predicted is there and is unlikely to be chance, but the size is negligible, and a university would not be able to detect it in the behaviour of its own alerts. This is a result worth stating plainly, because it runs against the expectation the portability literature creates: at least on this dataset, at this prediction day, a model does not need to have met a course to rank the students on it.

The exception is instructive. Module GGG is the hardest module to predict under every design: every figure it produces falls between 0.6364 and 0.7313, while no other module drops below 0.7442 and four of the seven never fall under 0.80. It also supplies two of the five presentations where withholding the course helped rather than hurt. GGG and AAA, the two lowest-scoring modules, are also the two where early unregistration is rare, at 3.9 and 4.1 percent against between 13.0 and 21.0 percent on the other five. The two facts are connected, and Section VI-F explains why.

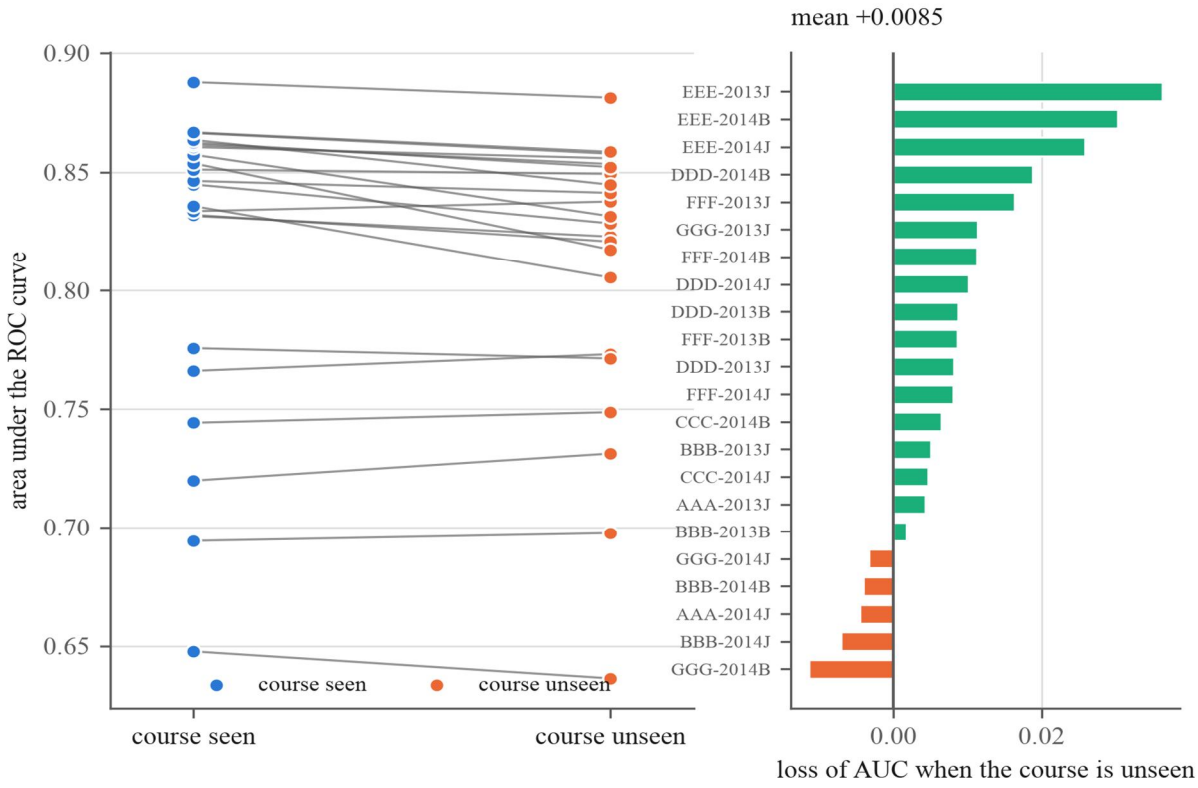


Fig. 2 The same students, scored by a model that has and has not seen their course. Each line is one of the twenty-two module-presentations.

E. Last year's model on this year's students

H2 asks the same question about time. Trained on the 13,529 registrations from 2013 and applied to the 19,064 from 2014, the model reaches 0.8241. Restricted to 2014 presentations of modules that also ran in 2013 it reaches 0.8128, and on CCC alone, a module absent from the training year entirely, it reaches 0.8606, which is higher rather than lower and is a further sign that novelty of course is not what limits these models.

The paired comparison is again on identical test rows. For each 2014 presentation, a model trained only on 2013 is set against a model trained on everything except that presentation. The means are 0.8131 and 0.8169, a difference of 0.0037. The Wilcoxon test gives W = 23.0, p = 0.12720, and the bootstrap interval on the difference runs from -0.0441 to 0.0520. H2 is not supported. A year of drift costs nothing that can be measured here, as Fig. 3 shows presentation by presentation.

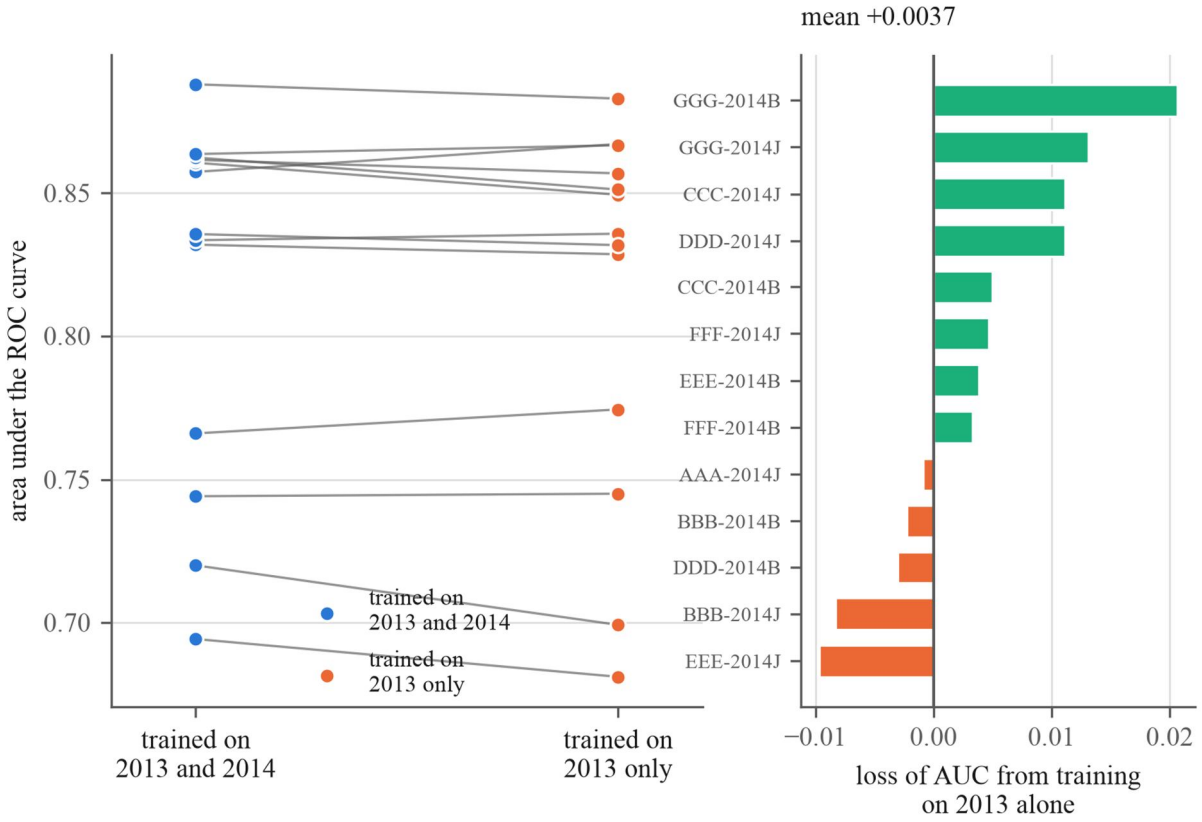


Fig. 3 Running last year's model on this year's students. Each line is one of the thirteen 2014 module-presentations.

F. Who the model is actually separating

The second part of H3 changes the population instead of the split. Table II gives the same random five-fold protocol applied to three nested groups of students.

TABLE II
WHAT THE POPULATION IS WORTH

Population	n	At-risk rate	AUC	95% CI	Recall at 20%
All registrations, as commonly modelled	32,593	0.528	0.8450	0.8405-0.8491	0.373
Still registered on day 30	27,466	0.440	0.7875	0.7819-0.7928	0.384
Reached the end of the module	22,437	0.314	0.7969	0.7910-0.8029	0.452

Random five-fold cross-validation, gradient-boosted model, identical features throughout. The third row excludes every withdrawal, so its outcome is failure among students who completed.

Removing the 5,127 students, 15.7 percent of the file, who had already unregistered before the prediction day costs 0.0575 of area under the curve. Measured above chance that is a sixth of everything the model appeared to achieve. Removing every withdrawal costs 0.0481, or 13.9 percent above chance. H3b is supported, and it is the largest single effect in this paper.

The reason is visible in the scores themselves. The median predicted risk is 0.196 for a student who finishes with a distinction, 0.323 for a pass, 0.595 for a fail and 0.926 for a withdrawal. Cliff's delta against students who passed is 0.760 for withdrawals and 0.590 for failures, both large, but the gap between those two numbers is the point. Considered as a detector of withdrawal the model scores 0.8799; considered as a detector of failure among students who stayed the course it scores 0.7948. Fig. 4 shows the four outcome groups side by side.

This is the finding with the sharpest practical edge. The system is best at recognising the student who has stopped opening the course website, and that student is often not a prediction at all but a record of something that has already happened. The student the intervention exists for, the one still turning up and quietly failing, is the one the model sees least well.

If that is right, then modules where hardly anyone leaves early should be the modules that measure worst, because they offer the model none of the easy cases. They are. Ordering the seven modules by the share of students who had unregistered by day 30 orders them almost exactly by measured accuracy: GGG at 3.9 percent departure scores 0.6873, AAA at 4.1 percent scores 0.7600, and CCC at 21.0 percent scores 0.8742, with only BBB out of place. The rank correlation across the seven is 0.750 ($p = 0.0522$) and the linear correlation is 0.876 ($p = 0.0098$). Seven modules is far too small a sample for this to stand on its own, and it was not pre-registered; it is reported because it points the same way as the population test above, not as evidence in its own right.

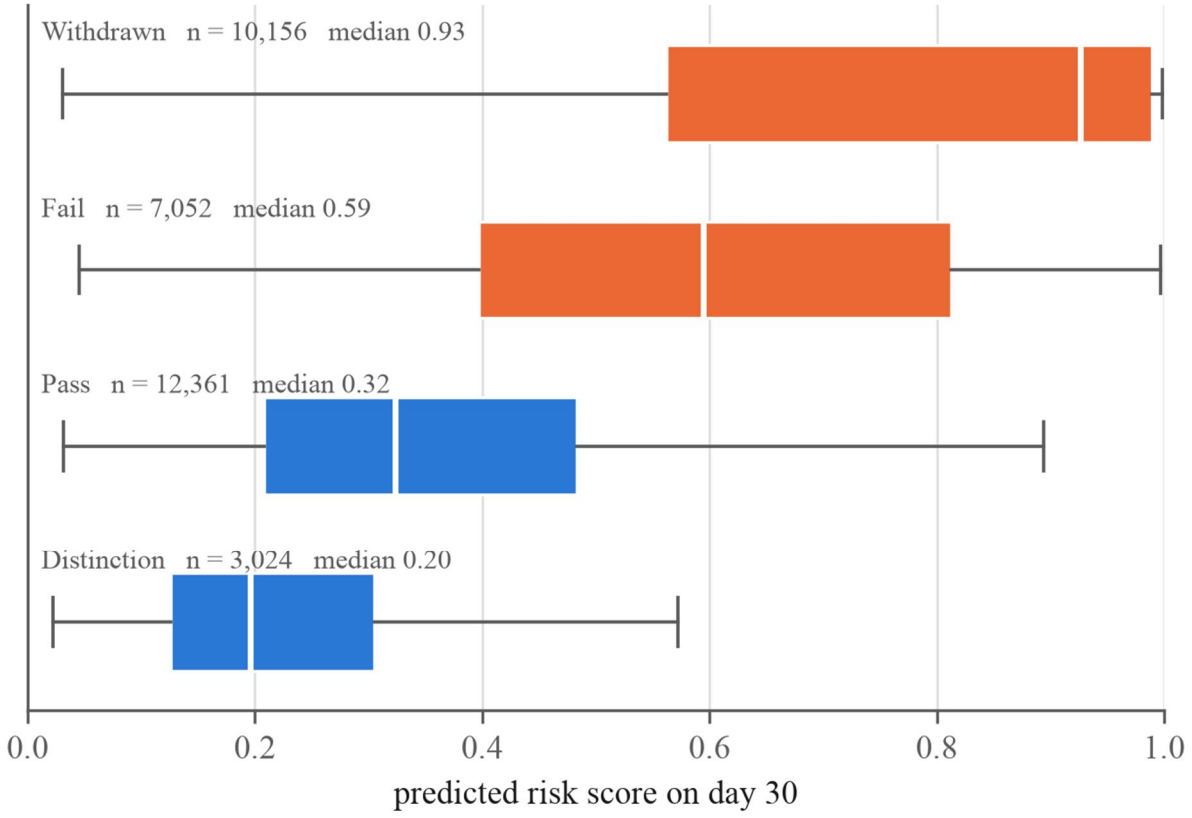


Fig. 4 What the model finds easy and what it does not.

G. One threshold, and who it misses

H4 moves from ranking to action. The 2013-trained model is applied to the 2014 cohort, and a single cut-off is fixed at the score that would flag the highest-scoring 20 percent of the training data. The cut-off is chosen before the test set is looked at, which is how a deployed system would set it. It flags 23.5 percent of the 2014 students.

At that cut-off the alert misses 58.6 percent of the students who went on to fail or withdraw. This figure deserves to sit beside the 0.8450 in any procurement conversation, because it is the same model measured in the currency of what happens to students.

The misses are not evenly spread, and not in the direction usually feared. Among at-risk students in the most deprived decile the alert missed 52.7 percent; among at-risk students in the least deprived decile it missed 64.6 percent. The difference is -0.118, with a 95 percent interval from -0.165 to -0.072, z = -4.87 and $p = 1.12 \times 10^{-6}$; the minus sign records the direction, which is that the most deprived band is the one missed less often. The odds of being missed for a most-deprived student are 0.613 times those of a least-deprived one, with an interval from 0.503 to 0.747. Across all ten bands the chi-square is 80.45 on 9 degrees of freedom, $p = 1.32 \times 10^{-13}$, and Cramer's V is 0.0896, which is negligible. H4 is supported in the sense that the miss rates genuinely differ, and unsupported in the sense that would matter to a fairness audit, since the association is weak and favours the more deprived group. Table III gives every band, and Fig. 5 plots them.

TABLE III
MISSED AT-RISK STUDENTS BY DEPRIVATION BAND, 2014 COHORT

Deprivation band	At-risk n	Missed	Miss rate	False alarm rate
0-10% (most deprived)	1,223	645	0.527	0.049
10-20%	1,268	670	0.528	0.056
20-30%	1,313	702	0.535	0.040
30-40%	1,115	657	0.589	0.024
40-50%	1,029	606	0.589	0.023
50-60%	956	577	0.604	0.035
60-70%	862	553	0.642	0.023
70-80%	833	527	0.633	0.013
80-90%	781	483	0.618	0.016
90-100% (least deprived)	632	408	0.646	0.008

Miss rate is the share of genuinely at-risk students the alert did not flag. False alarm rate is the share of students who were not at risk who were flagged anyway, computed on the separate population of not-at-risk students.

The final column explains the pattern and removes any comfort from it. Deprived students are flagged more often in general, because they fail more often in general, and a model that scores them higher will both catch more of the ones who are struggling and raise more false alarms about the ones who are not. The false-alarm rate runs from 0.049 in the most deprived decile to 0.008 in the least deprived, a sixfold difference in the opposite direction to the miss rate. This is the trade-off Kizilcec and Lee describe: when base rates differ, equal miss rates and equal false-alarm rates cannot both be had [14]. Neither column is an error. Which one a university should care about is a decision about what the alert is for, and it is not a decision a model can make.

Prior education shows the same shape more sharply. Among at-risk students with no formal qualifications the alert missed 40.9 percent; among those holding a postgraduate qualification it missed 73.3 percent. The difference is -0.324, interval -0.451 to -0.196, z = -4.58, $p = 4.67 \times 10^{-6}$, odds ratio 0.252 with an interval from 0.137 to 0.463. The chi-square across all five levels is 57.34 on 4 degrees of freedom, $p = 1.05 \times 10^{-11}$, with Cramer's V of 0.0748, again negligible. The postgraduate group is small, at 75 at-risk students, and the interval is correspondingly wide.

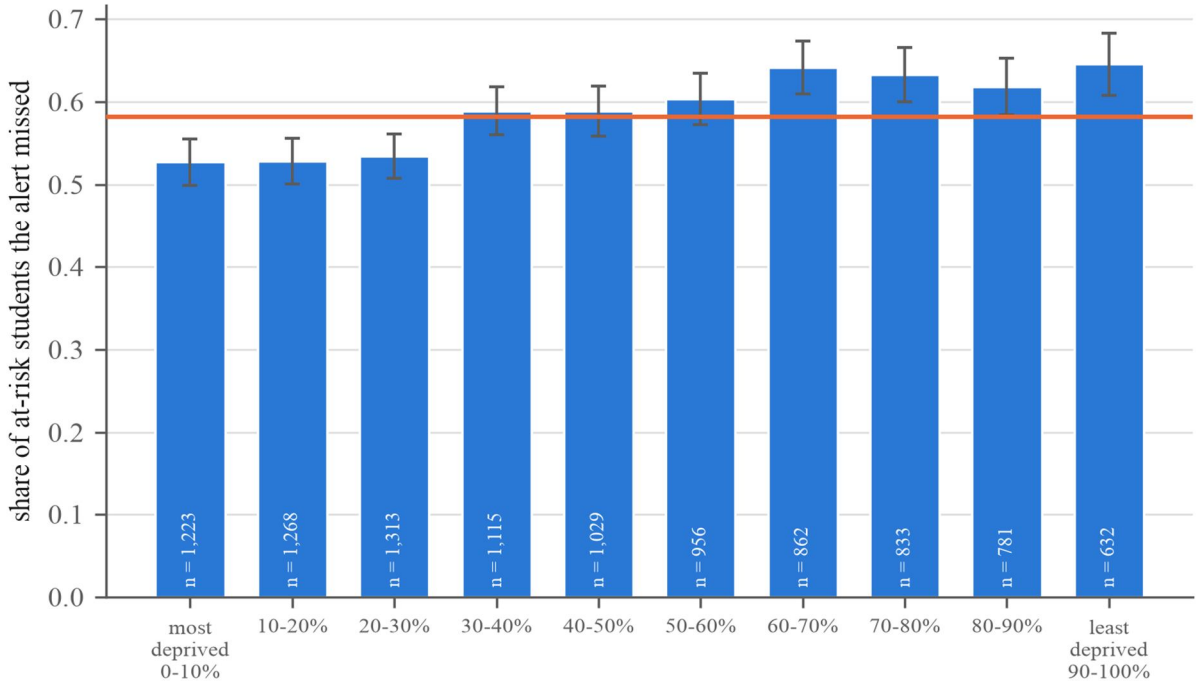


Fig. 5 Who the alert misses, by neighbourhood deprivation. The horizontal line is the miss rate across all at-risk students with a recorded band, 0.582. Bars carry 95 per cent intervals.

H. The same comparison inside each module

H5 asks whether the deprivation gap is a fact about the model or a fact about who enrols where. Recomputed within each module separately, the gap between the most and least deprived deciles is smaller in five of the six modules with enough students to test, and reverses in one. Table IV sets the within-module figures against the pooled one, and Fig. 6 plots the same comparison.

TABLE IV
THE DEPRIVATION GAP POOLED AND WITHIN MODULES

Module	Miss rate, most deprived	Miss rate, least deprived	Gap
BBB	0.433 (n = 328)	0.518 (n = 85)	-0.085
CCC	0.439 (n = 269)	0.650 (n = 206)	-0.212
DDD	0.470 (n = 168)	0.645 (n = 110)	-0.175
EEE	0.646 (n = 82)	0.706 (n = 51)	-0.060
FFF	0.555 (n = 272)	0.588 (n = 131)	-0.033
GGG	0.990 (n = 100)	0.941 (n = 34)	+0.049
Pooled	0.527 (n = 1,223)	0.646 (n = 632)	-0.118

A negative gap means the alert missed fewer of the most deprived students. Module AAA is not tested: it contributes only 4 and 15 at-risk students to the two bands.

The unweighted mean of the within-module gaps is -0.086 against a pooled gap of -0.118, so roughly a quarter of the pooled disparity is composition rather than model behaviour. The reason is visible in enrolment: the most deprived decile makes up 69.3 percent of the two-band population on BBB and 65.7 percent on GGG, but only 27.0 percent on AAA. Modules where deprived students concentrate are also modules where the alert behaves differently, and pooling mixes the two.

GGG reverses outright. There the alert missed 99.0 percent of at-risk students in the most deprived decile and 94.1 percent in the least, so the direction of the disparity flips, on a module where the alert is close to useless for everybody. H5 is therefore partly supported: five of six modules agree with the pooled direction, the pooled magnitude overstates the typical within-module magnitude, and one module contradicts it. A fairness audit run only on the pooled population would have reported a cleaner and more favourable story than the data contain.

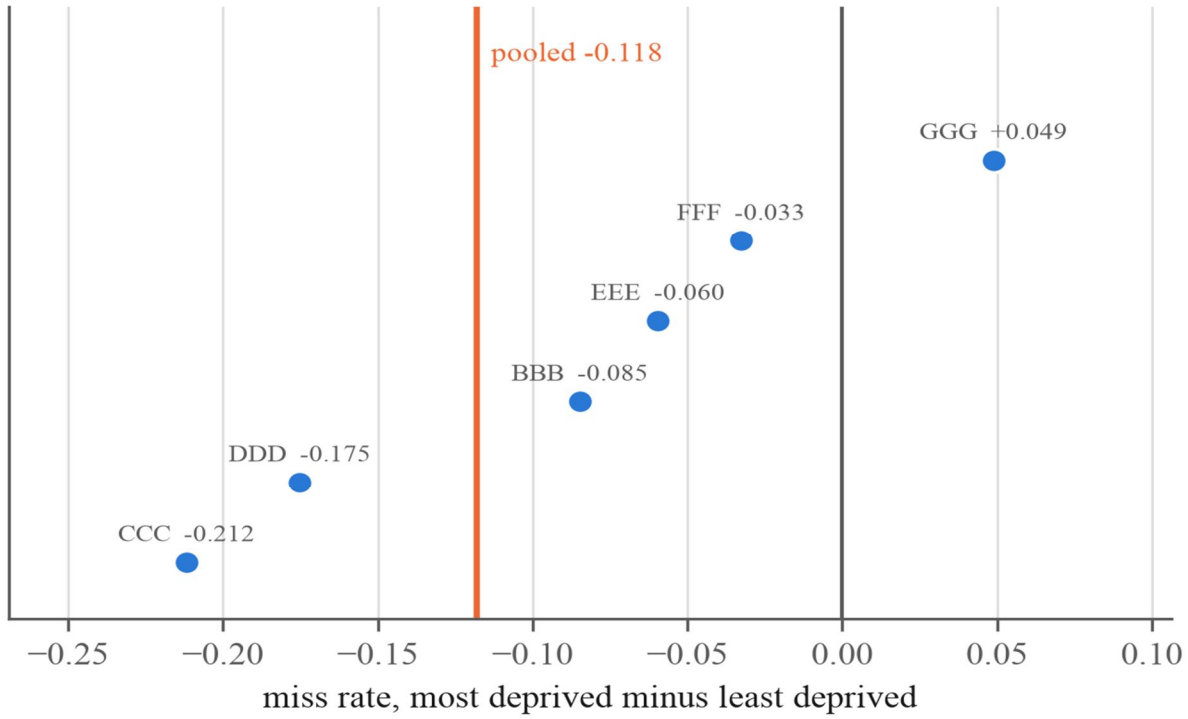


Fig. 6 The pooled gap and the gap inside each module.

I. What the model leans on

Permutation importance was measured on the 2013-trained model applied to the 2014 cohort, as the fall in area under the curve when one column is shuffled, averaged over five shuffles from a fixed seed against a baseline of 0.8241. The ranking is given in Table V.

TABLE V
WHAT SHUFFLING A COLUMN COSTS, 2014 COHORT

Column	Fall in AUC	Group
mean_score	0.0395	assessment
assessments_due_by_30	0.0299	assessment
days_since_last_click	0.0265	behaviour
min_score	0.0142	assessment
highest_education	0.0068	declared
studied_credits	0.0063	declared
clicks_wk2	0.0037	behaviour
imd_band	0.0033	declared
term	0.0032	declared
num_of_prev_attempts	0.0027	declared

Baseline area under the curve 0.8241. The ten largest of 39 columns.

Summed over all columns with a positive contribution, declared and registration attributes account for 16.5 percent of the total and behaviour and assessment for 83.5 percent. On this dataset, at this prediction day, that contradicts the reported 68 percent share for static demographic and pre-enrolment features [1] and agrees with the finding that the dominant signal is temporal and behavioural [6]. The disagreement is not necessarily a contradiction in the underlying data, since the three studies differ in prediction day, feature construction and importance method, but a reader choosing between them should know that the measurement here comes out firmly on one side.

The single most useful column is the mean mark on work already submitted, and the second is how many assessments the module had scheduled by day 30, which is a property of the course rather than of the student. That a course-level column ranks second among 39 is the same phenomenon as Section VI-C measured directly. Fig. 7 shows the twelve largest contributions.

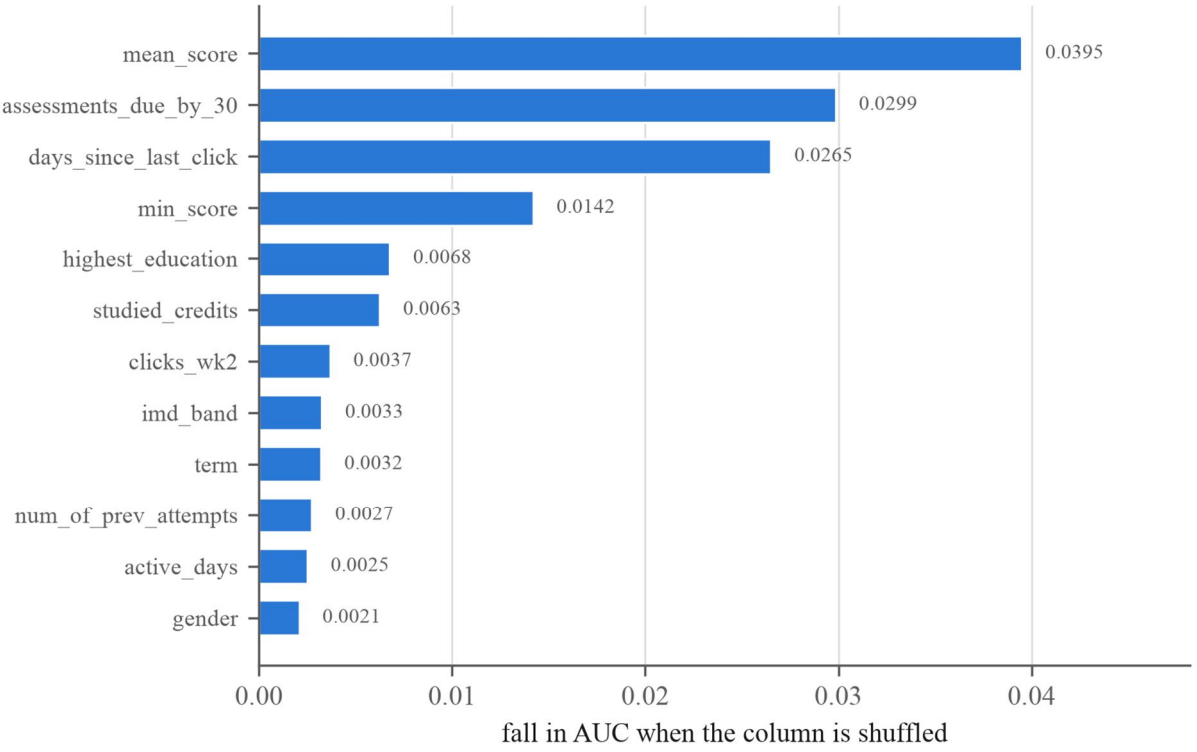


Fig. 7 What the model leans on, measured on the 2014 cohort.

J. The deployable figure, and where the headline goes

The three moves towards realism can be applied together. Each presentation is scored by a model that has never seen its module, and the test rows are then restricted to students who were still registered on day 30, since a student who has already unregistered cannot be helped by an alert issued afterwards. This is the closest the dataset allows to what a university would actually experience on the first day it switched such a system on.

Under that protocol the mean area under the curve across the twenty-two presentations is 0.7520, ranging from 0.6074 on GGG-2013J to 0.8271 on CCC-2014B, and the mean recall at a 20 percent alert budget is 0.360. Against the pooled random-split figure of 0.8450 that is a fall of 0.0931. Table VI attributes it.

TABLE VI
WHERE THE HEADLINE FIGURE GOES

Evaluation design	AUC	Cost of this step	Share of the total
Pooled random split, all students	0.8450	-	-
Mean within one presentation, course seen	0.8183	0.0267	28.7%
Mean within one presentation, course unseen	0.8099	0.0085	9.1%
The same, students already gone removed	0.7520	0.0579	62.2%

Every row after the first is an unweighted mean of per-presentation areas under the curve; the first is the single pooled figure a random split produces. The three costs sum to the 0.0931 difference between the first row and the last.

Almost two thirds of the distance between the reported figure and the deployable one is students who had already left. Something over a quarter is the arithmetic of pooling courses. Under a tenth is the model meeting a course it does not know, which is the only one of the three that is a machine learning problem at all.

VII. RESULTS

Of the five hypotheses written down in advance, two are supported, two are not, and one is supported in part.

1) H1, that an unseen course costs materially: NOT SUPPORTED. The cost is 0.0085 of area under the curve, consistent in direction across 17 of 22 presentations and significant at p = 0.00219, but negligible in size and invisible at an operational alert budget.
2) H2, that a year of drift costs materially: NOT SUPPORTED. The cost is 0.0037, p = 0.12720, with a bootstrap interval spanning zero.
3) H3, that much of the figure comes from course composition and from students who have already stopped: SUPPORTED, in both parts. Presentation identity alone reaches 0.5777, covering 22.5 percent of the model's above-chance discrimination; removing students who had already unregistered costs 0.0575.
4) H4, that misses are unevenly distributed: SUPPORTED as a difference and weak as an effect. The gap between the most and least deprived deciles is 0.118, p = 1.12 x 10-6, with Cramer's V of 0.0896 across all bands, and it favours the most deprived.
5) H5, that the disparity holds within modules: PARTLY SUPPORTED. Five of six modules agree in direction, the within-module mean gap of -0.086 is smaller than the pooled -0.118, and one module reverses.

Combining the realistic choices gives the study's headline number. Scored on an unseen module and restricted to students still enrolled on the prediction day, the mean area under the curve is 0.7520 and mean recall at a 20 percent alert budget is 0.360. The 0.0931 fall from the pooled figure divides 62.2 percent to departed students, 28.7 percent to pooling across courses and 9.1 percent to the course being unfamiliar.

VIII. DISCUSSION

A. *The inflation is in the test set, not the model*

The expectation this study began with was that these models would transfer badly, and that the gap between published accuracy and deployed accuracy would be a generalisation failure. That expectation was wrong in an interesting way. Generalisation is the smallest term in the decomposition. What separates the reported figure from the deployable one is, first, students who had already left, and second, the arithmetic of pooling courses that fail at very different rates.

Both of those are properties of how a test set is assembled, and neither is fixed by a better algorithm. A team that improves its model and reports the improvement under the same pooled protocol will keep both sources of inflation intact. This reframes what the portability literature has been asking for. Gasevic and colleagues were right that course context matters [10], but on this evidence it matters mainly by inflating pooled evaluation, not by breaking transfer.

B. *What an alert is for*

The most uncomfortable result is that the model is substantially a withdrawal detector. Withdrawal is often recorded when a student unregisters, and by that point the decision is made. Ranking those students highly is easy, adds a sixth of the headline discrimination, and helps nobody. Meanwhile the student who is quietly failing while still attending is the harder case, and is the case the intervention was designed for. This suggests that at-risk should not be one label. A system that reported two scores, one for disengagement already underway and one for likely failure among students still engaged, would be less impressive on a single metric and more useful in an office. It would also stop the easy population from subsidising the hard one in the published number.

C. *Fairness that runs the other way*

The equity result is not the one the literature primes a reader to expect. The alert catches more of the at-risk students in deprived neighbourhoods than in affluent ones, and raises far more false alarms about deprived students who were going to be fine. Both follow from the same mechanism, the higher base rate, and the impossibility Kizilcec and Lee describe means no threshold removes the trade-off [14]. Bird, Castleman and Song make the connected point that the harm, if there is one, appears where the threshold allocates support [18]. Here the allocation is skewed towards deprived students in both a helpful and an unhelpful sense at once, and a university would have to decide which it minds. The within-module check matters for that decision, because a quarter of the apparent advantage turns out to be composition, and on one module the sign flips.

D. What this means for reading a published figure

The practical translation is a discount. A reader meeting an area under the curve in an at-risk paper should ask three questions before believing it applies to their institution: how many courses were pooled to produce it and how much their failure rates differ; whether students who had already withdrawn were in the test set; and what recall the model achieves at the number of students the institution can actually contact. On this dataset those three questions account for essentially the whole difference between the reported figure and the deployable one, and none of them is about the model.

IX. LIMITATIONS AND THREATS TO VALIDITY

A. One institution, one decade

This is a study of seven modules at one distance-teaching university in 2013 and 2014. The Open University is unusual: it has open entry, its students are older and more likely to be studying part-time, and almost all teaching is mediated by the virtual learning environment, so the behavioural record is unusually complete. A campus university with lecture attendance and library visits would produce a different picture, and probably a weaker behavioural signal. Nothing here should be read as a general law about early-warning systems.

The data are also more than a decade old, which matters most for the sections that are about the technology being sold today. What is measured here is a property of evaluation design rather than of any particular model, and evaluation designs have not changed, but a modern virtual learning environment logs different things and a modern cohort behaves differently.

B. Choices that were fixed but could have been otherwise

Day 30 was chosen in advance and not revisited, but it is one point on a curve. A prediction made at day 60 would have more assessment evidence and less time to act on it, and the balance between the terms in the decomposition would shift. The 20 percent alert budget is likewise a stand-in for a real institution's capacity, which nobody publishes. The at-risk definition follows the convention in this literature of merging failure with withdrawal, and Section VI-F is in effect an argument that the convention is a mistake; a study that had split the outcome from the start would have produced a cleaner analysis and a less direct comparison with existing work.

C. What the comparisons can and cannot show

The seen-versus-unseen contrast is about familiarity with a module, not about chronology: in some pairs the model that has seen the course has seen a later presentation of it. That is deliberate, since the aim was to isolate course familiarity, and the forward-in-time comparison handles chronology separately. But it means the two designs are not simply harder and easier versions of one thing.

The forward-in-time test rests on a single boundary, 2013 to 2014, with seven modules. It cannot distinguish a year in which nothing drifted from a general absence of drift. With thirteen 2014 presentations the paired test has limited power, and a null result is a failure to detect rather than a demonstration of absence.

Permutation importance is a statement about one fitted model on one test set, not about the underlying causes of failure. Correlated columns share credit and can each look unimportant. The comparison with the two published importance claims [1], [6] is therefore an indication, not an adjudication.

D. Statistical caution

With 32,593 rows almost any difference will reach significance, which is why every test here is paired with an effect size [22]. Two of the paper's results are significant and negligible at the same time, and the negligible half is the half that matters. The presentation-level bootstrap intervals are wider than the row-level ones by roughly a factor of three, and the wider ones should be believed. The deprivation comparisons in Section VI-H rest on small counts in some modules, and module AAA could not be tested at all.

Finally, this study observes; it does not intervene. Nothing here shows that contacting a flagged student changes their outcome, which is the question a university actually wants answered and which no observational dataset can settle.

X. RECOMMENDATIONS AND FUTURE WORK

Four recommendations follow directly from the measurements above, and are aimed at anyone publishing an at-risk model.

Report the within-course figure alongside the pooled one. If a paper pools courses, it should say how many and give the spread of their failure rates, and report the mean area under the curve within a course as well as across the pooled sample. On this dataset those two figures differ by 0.0267.

State what happened to students who had already left. Whether students who unregistered before the prediction date are in the test set changes the headline by more than anything else measured here. It should be a line in the methods, not an inference from the row count.

Report recall at a stated alert budget. An area under the curve does not tell a support team how many students they will reach. The same model that scores 0.8450 misses 58.6 percent of at-risk students when 20 percent of the cohort can be contacted.

Run the fairness audit within courses as well as across them. A quarter of the pooled deprivation gap here is composition, and one module reverses the sign.

Three extensions would strengthen the case. The first is to repeat the decomposition across the prediction horizon rather than at a single day, which would show whether the withdrawal term shrinks as genuine failure becomes predictable. The second is to model failure and disengagement as separate outcomes and report both, which Section VI-F argues is the more honest framing. The third is to replicate this on a contemporary dataset from a campus institution, where the behavioural record is thinner and the terms may rank differently.

XI. CONCLUSION

An at-risk model built on the Open University Learning Analytics Dataset and evaluated the way this literature evaluates such models reaches an area under the ROC curve of 0.8450 on day 30 of a course. Taking that figure apart shows that the part of it a university would actually experience is smaller, and that the shortfall is not where it was expected to be.

Model transfer, the failure mode the portability literature warns about, turns out to be cheap: an unfamiliar course costs 0.0085 of area under the curve and a year of drift costs 0.0037, neither of which a support office would notice. The expensive terms are structural. Pooling students from courses that fail at rates between 27.4 and 65.8 percent contributes discrimination that no single course contains, to the point that a model told only which presentation a student is on reaches 0.5777 while knowing nothing about the student. And 15.7 percent of the students in the file had already unregistered before the prediction was made; removing them costs 0.0575, the largest single term, because the model is markedly better at recognising a student who has already gone than one who is still present and failing.

Put together, a model scoring students on a module it has never seen, among only those still enrolled on the day the prediction is made, averages 0.7520. The 0.0931 that separates that from the published-style figure is 62.2 percent departed students, 28.7 percent pooled courses and 9.1 percent unfamiliarity.

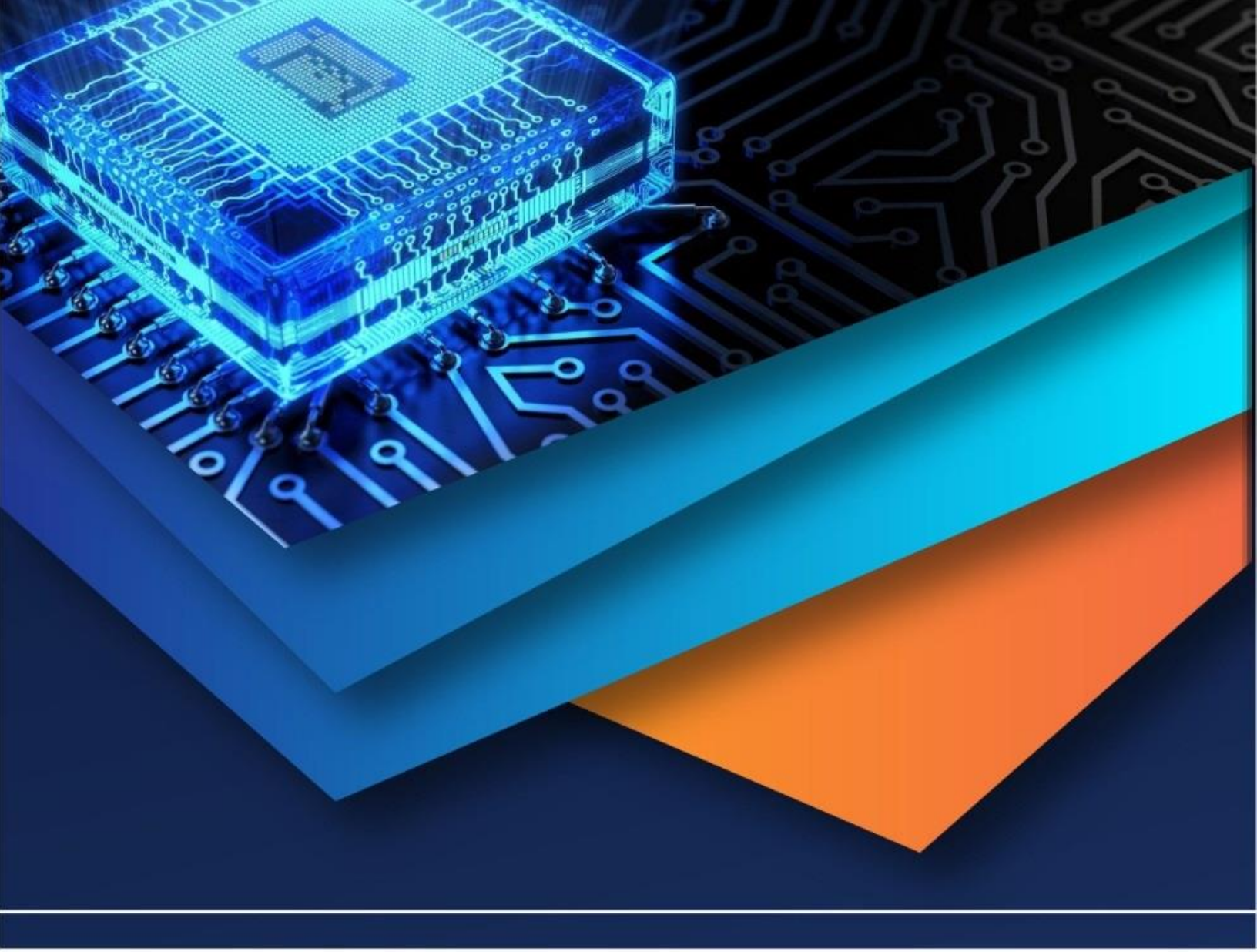
At an alert budget of one student in five, the model misses 58.6 percent of those who fail or withdraw. It misses the least deprived more often than the most deprived, and raises six times as many false alarms about the most deprived, both consequences of the same difference in base rates; a quarter of that pooled gap dissolves when deprivation is compared inside courses rather than across them, and on one module it reverses.

None of this means early-warning systems should not be built. It means the number on the brochure is a property of an evaluation protocol, and that a university reading it is entitled to ask which students were in the test set, how many courses were stirred together to produce it, and how many people the institution can actually telephone.

REFERENCES

[1] V. Kaushal and R. Mall, "AI-driven early warning systems for student success: discovering static feature dominance in temporal prediction models," arXiv:2512.12493, 2025. doi: 10.48550/arXiv.2512.12493.
[2] M. El Jihaoui, O. E. K. Abra, and K. Mansouri, "Towards a fair and transparent early warning system for identifying students at risk of academic failure," Frontiers in Education, vol. 11, art. 1878686, 2026. doi: 10.3389/feduc.2026.1878686.
[3] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," Scientific Data, vol. 4, art. 170171, 2017. doi: 10.1038/sdata.2017.171.
[4] E. Howard, "ouladFormat R package: preparing the Open University Learning Analytics Dataset for analysis," arXiv:2501.08366, 2025. doi: 10.48550/arXiv.2501.08366.
[5] M. Hlosta, D. Herrmannova, L. Vachova, J. Kuzilek, Z. Zdrahal, and A. Wolff, "Modelling student online behaviour in a virtual learning environment," arXiv:1811.06369, 2018. doi: 10.48550/arXiv.1811.06369.
[6] R. da Silva, J. Eicher, and G. Longo, "A unified survival benchmark for temporal dropout risk prediction in learning analytics," arXiv:2604.08870, 2026. doi: 10.48550/arXiv.2604.08870.
[7] C. Labba and A. Boyer, "Towards an online incremental approach to predict students performance," arXiv:2407.10256, 2024. doi: 10.48550/arXiv.2407.10256.
[8] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," arXiv:1606.06364, 2016. doi: 10.48550/arXiv.1606.06364.
[9] J. Gardner and C. Brooks, "Student success prediction in MOOCs," User Modeling and User-Adapted Interaction, vol. 28, no. 2, pp. 127-203, 2018. doi: 10.1007/s11257-018-9203-z.

[10] D. Gasevic, S. Dawson, T. Rogers, and D. Gasevic, "Learning analytics should not promote one size fits all: the effects of instructional conditions in predicting academic success," The Internet and Higher Education, vol. 28, pp. 68-84, 2016. doi: 10.1016/j.iheduc.2015.10.002.
[11] V. Swamy, M. Marras, and T. Kaser, "Meta transfer learning for early success prediction in MOOCs," Proc. 9th ACM Conf. on Learning @ Scale, pp. 121-132, 2022. doi: 10.1145/3491140.3528273.
[12] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, and L. A. Celi, "The myth of generalisability in clinical research and machine learning in health care," The Lancet Digital Health, vol. 2, no. 9, pp. e489-e492, 2020. doi: 10.1016/S2589-7500(20)30186-2.
[13] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," Nature Machine Intelligence, vol. 3, no. 3, pp. 199-217, 2021. doi: 10.1038/s42256-021-00307-0.
[14] R. F. Kizilcec and H. Lee, "Algorithmic fairness in education," arXiv:2007.05443, 2020. doi: 10.48550/arXiv.2007.05443.
[15] R. Yu, H. Lee, and R. F. Kizilcec, "Should college dropout prediction models include protected attributes?," Proc. 8th ACM Conf. on Learning @ Scale, pp. 91-100, 2021. doi: 10.1145/3430895.3460139.
[16] S. Riazy, K. Simbeck, and V. Schreck, "Fairness in learning analytics: student at-risk prediction in virtual learning environments," Proc. 12th Int. Conf. on Computer Supported Education, pp. 15-25, 2020. doi: 10.5220/0009324100150025.
[17] H. Anahideh, N. Nezami, P. Haghighat, and D. Gandara, "Auditing the imputation effect on fairness of predictive analytics in higher education," arXiv:2109.07908, 2021. doi: 10.48550/arXiv.2109.07908.
[18] K. A. Bird, B. L. Castleman, and Y. Song, "Are algorithms biased in education? Exploring racial bias in predicting community college student success," Annenberg Institute at Brown University, EdWorkingPaper 23-717; published version in the Journal of Policy Analysis and Management, vol. 44, pp. 379-402, 2024, 2023. doi: 10.26300/yd7z-6e20.
[19] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauria, J. R. Regan, and J. D. Baron, "Early alert of academically at-risk students: an open source analytics initiative," Journal of Learning Analytics, vol. 1, no. 1, pp. 6-47, 2014. doi: 10.18608/jla.2014.11.3.
[20] R. S. Baker, N. Bosch, S. Hutt, A. F. Zambrano, and A. J. Bowers, "On fixing the right problems in predictive analytics: AUC is not the problem," arXiv:2404.06989, 2024. doi: 10.48550/arXiv.2404.06989.
[21] O. Zawacki-Richter, V. I. Marin, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education - where are the educators?," International Journal of Educational Technology in Higher Education, vol. 16, art. 39, 2019. doi: 10.1186/s41239-019-0171-0.
[22] D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs," Frontiers in Psychology, vol. 4, art. 863, 2013. doi: 10.3389/fpsyg.2013.00863.
[23] R. A. Kievit, W. E. Frankenhuis, L. J. Waldorp, and D. Borsboom, "Simpson's paradox in psychological science: a practical guide," Frontiers in Psychology, vol. 4, art. 513, 2013. doi: 10.3389/fpsyg.2013.00513.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY