



PENALIZED LOGISTIC VERSUS ADDITIVE SPLINE MODELS FOR PREDICTING ACADEMIC FAILURE: A REPEATED NESTED CROSS-VALIDATION STUDY OF DISCRIMINATION AND CALIBRATION

Anju J. B.¹, Alby Joseph²

¹Assistant Professor in Statistics, Department of Statistics, Newman College, Thodupuzha, Kerala, India.

E-mail: anjudwaraka71@gmail.com, ORCID: 0002-0845-6921

²Assistant Professor in Statistics, Department of Statistics, Newman College, Thodupuzha, Kerala, India.

E-mail: josepha545@gmail.com, ORCID: 0003-5030-8580

Corresponding author: Dr. Anju J.B., anjudwaraka71@gmail.com

ARTICLE INFO

Received: 19/06/2026

Revised: 22/09/2026

Accepted: 11/10/2026

Published: 26/10/2026

DOI: <https://doi.org/10.69980/d7xydk10>

Volume:12, Issue:04, October 2026

Abstract

Statistical models that estimate a student's risk of an unsuccessful academic outcome are increasingly proposed for allocating academic support, but such models must both rank students accurately and produce well-calibrated probabilities. This study compared ridge-penalized logistic regression with additive logistic models built from cubic B-spline bases for predicting academic failure, defined as failure or withdrawal. Five nested specifications added background, engagement and assessment predictors in sequence and were evaluated by stratified five-fold cross-validation repeated three times, with the penalty tuned in an inner loop. Performance was measured by the area under the receiver operating characteristic curve (AUC), average precision, Brier score, log loss and calibration intercept and slope. Background variables alone gave no improvement in Brier score over a constant-risk prediction (AUC 0.543; Brier score 0.1254). Engagement variables added little (AUC 0.594; Brier score 0.1240), whereas assessment variables produced the largest gain: the assessment-inclusive logistic model achieved an AUC of 0.736, average precision 0.318, Brier score 0.1144 and calibration slope 0.942. Its spline counterpart did not improve on it (AUC 0.721; Brier score 0.1154; slope 0.904). With 147 unsuccessful outcomes, added flexibility did not improve internal predictive performance, favouring the parsimonious logistic model. The results are internal to one district survey and require external and temporal validation before any operational use.

Keywords: academic failure; additive models; calibration; learning analytics; nested cross-validation; penalized logistic regression; regression splines.

1. Introduction

1.1 Background

Academic progression and completion are central concerns for educational institutions. Learning management systems, also called virtual learning environments (VLEs), now record a wide range of digital traces, including visits to course resources, quiz activity, participation in discussion forums and the submission of assessments. These traces make it possible to build statistical models that describe study behaviour and estimate each student's probability of an unsuccessful outcome. Educational data mining and learning analytics developed partly in response to this availability of fine-grained educational data (Baker & Yacef, 2009; Romero & Ventura, 2007, 2020).

Prediction is only one aspect of institutional decision-making. When risk scores are used to identify students who may need additional support, models must demonstrate both good discrimination and adequate calibration. Discrimination refers to a model's ability to distinguish between students who experience academic failure and those who do not, whereas calibration indicates how closely predicted probabilities correspond to the observed frequency of failure. These properties are distinct: a model may achieve an area under the receiver operating characteristic curve (AUC) greater than 0.5 while systematically overestimating or underestimating the probability of failure for certain students (Van Calster et al., 2019). Since predicted probabilities may inform decisions about which students receive support and how institutional resources are allocated, calibration is as important as discrimination. Furthermore, distinguishing between explaining relationships among variables and predicting outcomes for new observations is essential to defining the analytical objective clearly (Shmueli, 2010).

Educational theory informs the interpretation of engagement variables. Tinto (1975) linked persistence to students' academic and social integration, and Fredricks et al. (2004) described engagement as having behavioural, emotional and cognitive dimensions. Self-regulated learning frameworks emphasize planning, monitoring and the strategic use of learning resources (Pintrich, 2004; Zimmerman, 2002). These perspectives make activity measures plausible predictors, but they do not imply that a click count measures learning directly or that a predicted probability represents a causal effect.

Learning analytics also raises questions about the meaning and governance of student data. Ethical principles include transparency about data provenance, proportionality of data collection and students' opportunity to understand or contest automated inferences (Pardo & Siemens, 2014; Slade & Prinsloo, 2013). Theory is needed to interpret what an activity measure represents, rather than allowing a large database to define learning by itself (Wise & Shaffer, 2015). Research on student-facing dashboards and recommender systems further shows that predictions are embedded in social and pedagogical settings in which educators and learners may act on them (Bodily & Verbert, 2017).

Early-alert projects illustrate how a statistical model becomes part of an institutional workflow. Arnold and Pistilli (2012) described Course Signals at Purdue University as a learning analytics system intended to increase student success, and Jayaprakash et al. (2014) reported an open-source analytics initiative for the early identification of academically at-risk students. These examples show that a prediction requires a defined action, a communication process and an evaluation of consequences. The present study is confined to the model-comparison stage, because the survey did not include an intervention protocol or an operational decision threshold.

Against this background, the present study uses an online survey of students in Thiruvananthapuram (Trivandrum) district, Kerala, India, to examine how much background, engagement and assessment information contribute to predicting failure or withdrawal, and whether an additive spline form improves prediction over a linear logistic form once the information set is held fixed.

1.2 Review of literature

1.2.1 Educational data mining and learning analytics

Romero and Ventura (2007) reviewed educational data mining (EDM) studies published between 1995 and 2005, highlighting the use of diverse analytical techniques, including statistical analysis, visualization, clustering, classification, association analysis, and text mining. Their review demonstrated that educational data vary in structure and purpose, emphasizing the need to select analytical methods according to the research question and the nature of the data. Similarly, Baker and Yacef (2009) examined the early development of educational data mining and identified its broad range of applications, from predicting academic performance to investigating learner behaviour. They also highlighted the growing emphasis on predictive modelling and the discovery of meaningful patterns in educational data. Siemens and Baker (2012) explored the relationship between learning analytics and educational data mining, emphasizing that greater collaboration between the two communities could strengthen methodological development and improve the interpretation of educational findings. Subsequently, Romero and Ventura (2020) revisited the development of the field, documenting its expansion across academic, teaching, and institutional applications. They also highlighted reproducibility and the educational interpretation of analytical results as important directions for future research.

Learning analytics has also been framed as an institutional and disciplinary development. Ferguson (2012) identified the technological, educational and policy drivers of learning analytics, together with challenges in institutional capability, data access and the translation of analytics into practice. Siemens (2013) presented learning analytics as an emerging discipline drawing on several research traditions, emphasized the need to represent the complexity of learning processes more adequately and highlighted privacy, data ownership and institutional capacity. Viberg et al. (2018) examined 252 learning analytics papers in higher education and found that descriptive, interpretive and experimental approaches coexisted with varying levels of evidence; the label "learning analytics" therefore covers different research designs, and prediction should not be confused with demonstrated educational benefit.

1.2.2 Learning theory, context and the meaning of activity data

Gašević et al. (2015) argued that learning analytics should remain about learning rather than become a purely technical exercise in predicting observable behaviour, and called for stronger links between analytics, pedagogy and learning theory. Gašević et al. (2016) showed that predictive relationships change across instructional conditions and warned against one-size-fits-all interpretations: course design and context can moderate the value of a behavioural indicator. Conijn et al. (2017) compared performance prediction across 17 blended courses using Moodle data and found that results varied strongly across courses, so that the portability of prediction models was low; they also reported that LMS data were of little additional value once in-between assessment grades were taken into account. You (2016) examined behavioural indicators in learning management system data from an online course and found that regular study, assignment submission patterns, session frequency and evidence of reading were informative for course achievement, suggesting that measures related to meaningful learning behaviours can be more informative than simple frequency counts. Tempelaar et al. (2015) studied learning analytics in a data-rich context and asked which available information is most informative for feedback generation, illustrating that the usefulness of a variable depends on the feedback purpose and on how the variable is measured.

1.2.3 Prediction of academic performance and early warning

Macfadyen and Dawson (2010) used learning management system data to develop a proof-of-concept early-warning system for educators, demonstrating the practical appeal of combining activity indicators with academic information while showing that the design of an alert system involves choices about variables, timing and action. Akçapınar et al. (2019) evaluated an early-warning setting with interaction data, comparing algorithms and preprocessing choices at several points during a course; their findings illustrate the value of asking how early a prediction can be made, rather than reporting only an end-of-course result. Kuzilek et al. (2017) introduced the Open University Learning Analytics Dataset, which combines demographic variables, assessment results and daily summaries of virtual learning environment interactions; because its provenance and structure are documented, it has become a useful benchmark. The survey analysed in the present study uses similar variable categories but was collected locally from respondents in Thiruvananthapuram district; it is not a reproduction of the Open University data.

Several systematic reviews have synthesized this literature. Hellas et al. (2018) reviewed 357 studies of academic performance prediction and reported increasing methodological variety alongside concerns about research quality, feature selection and evaluation practices, cautioning against interpreting a single accuracy value as evidence of a generally useful model. Alyahyan and Düşteğör (2020) synthesized research on predicting academic success and offered a step-by-step guide to defining success, selecting attributes and choosing a model, emphasizing the consequences of the analyst's decisions. Namoun and Alshanqiti (2021) systematically reviewed 62 studies published between 2010 and November 2020; regression and supervised learning were frequently used, online activity, assessment grades and academic emotions were common predictors, and defining outcomes and comparing methods remained challenging.

1.2.4 From prediction to educational benefit

Ifenthaler and Yau (2020) reviewed evidence on using learning analytics to support study success in higher education and concluded that, although many approaches appeared promising, rigorous large-scale evidence of effectiveness remained limited. Alfredo et al. (2024) reviewed human-centred learning analytics and artificial intelligence in education and highlighted end-user involvement, human control, reliability and trustworthiness, placing technical performance within the relationships among students, educators and institutions. Together with the ethical frameworks discussed above, these reviews indicate that predictive performance is a necessary but not a sufficient condition for the responsible use of a risk score.

1.3 Research gap and contribution

Three gaps motivate this study. First, reviews of academic performance prediction have raised concerns about evaluation practice, including reliance on single accuracy measures (Hellas et al., 2018) and difficulties in defining outcomes and comparing methods (Namoun & Alshanqiti, 2021). Methodological work on clinical prediction models, by contrast, treats calibration and proper scoring rules as essential when predicted risks guide the allocation of resources (Steyerberg et al., 2010; Van Calster et al., 2016, 2019), and the TRIPOD reporting guideline asks for model performance to be reported in full (Collins et al., 2015). Second, when a flexible model is compared with a simpler model that uses a different set of predictors, gains from additional information cannot be separated from gains due to functional form. A design that adds predictor blocks in sequence and compares functional forms within each information set separates these two sources. Third, with a binary outcome of modest prevalence, the number of events limits how many parameters can be estimated reliably (Riley et al., 2020; van Smeden et al., 2016, 2019); this constraint bears directly on whether spline flexibility can be expected to improve prediction. In addition, the primary studies reviewed above were conducted outside India.

The study makes three contributions. First, it adopts a block-wise design in which background, engagement and assessment information are introduced sequentially. Within each information block, linear and additive model specifications are compared to distinguish the contribution of additional information from the benefits of greater model flexibility. Second, all model values are evaluated using discrimination, proper scoring rules and calibration within a repeated nested cross-validation framework. Identical outer folds are used across models to ensure fair comparisons and reduce optimistic bias arising when model tuning and performance evaluation use the same data (Cawley & Talbot,

2010; Varma & Simon, 2006). Third, it provides evidence on whether spline flexibility is beneficial when the number of unsuccessful outcomes is limited, using survey data from a district in Kerala, India. From a statistical perspective, the study demonstrates the importance of complementing ranking-based measures with proper scoring rules and calibration when comparing predictive models.

1.4 Objectives

The specific objectives of the study are:

- (i) to quantify the incremental predictive value of background, engagement and assessment predictor blocks for the composite outcome of failure or withdrawal;
- (ii) to compare ridge-penalized logistic regression with additive spline logistic models while holding the information set fixed;
- (iii) to evaluate and compare the discrimination, probabilistic accuracy and calibration of the models using repeated nested stratified cross-validation; and
- (iv) to describe the fitted relationships between selected engagement and assessment variables and predicted risk.

The analysis focuses on an internal comparison of candidate prediction models. It does not aim to estimate causal effects, assess the effectiveness of early-warning system or substitute for external validation. Section 2 describes the data, modelling approaches and validation design; Section 3 reports the results; Section 4 discusses the findings, limitations and directions for future research; and Section 5 concludes.

2 Materials and Methods

2.1 Study design and data collection

This study used a cross-sectional online survey design. Data were collected from 1 to 31 March 2026 through an online questionnaire distributed by e-mail (Gmail) to students recruited through schools and colleges in Thiruvananthapuram district, Kerala, India. The final dataset contained 1,000 completed responses on 14 variables, with one row per respondent. The questionnaire recorded background characteristics, learning-activity measures and assessment information. Ethical approval and consent procedures are reported in the Declarations.

2.2 Outcome

The outcome was each respondent's final academic result, recorded as Pass ($n = 742$), Distinction ($n = 111$), Fail ($n = 106$) or Withdraw ($n = 41$). The binary outcome was defined as $Y = 1$ for Fail or Withdraw and $Y = 0$ for Pass or Distinction, that is, $Y = 1 - \text{academic success}$. The unsuccessful group comprised 147 respondents (14.7%), of whom 106 (72.1%) failed and 41 (27.9%) withdrew. "Academic failure" is used as a concise label for this composite outcome; the separate counts are reported because failure and withdrawal may arise through different educational mechanisms.

2.3 Predictors

Ten predictors were grouped into three blocks (Table 1). The background block contained age band, highest education, studied credits and number of previous attempts; the engagement block contained the number of active VLE days, total VLE clicks, quiz clicks and forum clicks; and the assessment block contained the number of assessment attempts and the average assessment score. Age band (recorded as 0–35, 35–55 and 55 years or above) and highest education were treated as categorical variables; all other predictors were numerical. The three click counts were transformed before modelling. Resource clicks were recorded and are summarized descriptively in Section 3.1 but were not used as a predictor.

Table 1. Predictor blocks used in the study

Block	Variables	Type
Background	Age band; highest education	Categorical
	Studied credits; previous attempts	Numerical
Engagement	Active VLE days; total VLE clicks; quiz clicks; forum clicks	Numerical (click counts transformed)
Assessment	Assessment attempts; average assessment score	Numerical

Note. VLE, virtual learning environment. Resource clicks were recorded but not used as a predictor.

2.4 Statistical models

2.4.1 Binary response framework

Let Y_i denote the outcome of respondent i ($i = 1, \dots, n$; $n = 1,000$), x_i the corresponding vector of predictors and $p_i = P(Y_i = 1 | x_i)$. All models assume

$$Y_i | x_i \sim \text{Bernoulli}(p_i), \quad \text{logit}(p_i) = \log \left\{ \frac{p_i}{1-p_i} \right\} = \eta_i \quad (1)$$

where the linear predictor η_i is linear in the logistic specifications and additive in the spline specifications. This formulation follows the generalized linear model treatment of binary responses (Nelder & Wedderburn, 1972) and the regression analysis of binary sequences (Cox, 1958). Because the data contain no repeated observations, course identifiers or observation dates, the models are cross-sectional.

2.4.2 Ridge-penalized logistic regression

In the logistic specifications,

$$\eta_i = \beta_0 + \sum_{j=1}^q \beta_j x_{ij} \quad (2)$$

where β_0 is the intercept and β_j is the coefficient of the j th of q predictor columns. Numerical predictors entered linearly on the log-odds scale after transformation, and categorical predictors were represented by reference-coded indicator variables. Coefficients were estimated by minimizing the L2-penalized (ridge) negative log-likelihood (Hoerl & Kennard, 1970), written here in terms of the inverse penalty parameter C :

$$\hat{\beta}(C) = \arg \min_{\beta} \left[- \sum_{i=1}^n \{y_i \log p_i(\beta) + (1-y_i) \log (1-p_i(\beta))\} + \frac{1}{2C} \sum_{j=1}^q \beta_j^2 \right] \quad (3)$$

in which the intercept is not penalized and smaller values of C imply stronger shrinkage. Numerical predictor columns and spline basis columns were standardized using training-set means and standard deviations; indicator variables were not standardized. The value of C was selected from $\{0.01, 0.1, 1, 10\}$ by inner cross-validation (Section 2.5). Lasso and elastic-net penalties (Tibshirani, 1996; Zou & Hastie, 2005), for which coordinate-descent regularization paths provide an efficient computational approach (Friedman et al., 2010), are alternatives that were not fitted here.

2.4.3 Additive spline logistic models

In the additive specifications, selected numerical predictors entered through univariate spline functions:

$$\eta_i = \beta_0 + \mathbf{z}_i^T \gamma + \sum_{j \in S} f_j(x_{ij}), \quad f_j(x) = \sum_{k=1}^K \theta_{jk} B_{jk}(x) \quad (4)$$

where $\mathbf{z}_i$ contains the indicator variables and the numerical predictors retained as linear terms, S is the set of smoothed predictors and B_{jk} are cubic B-spline basis functions (Eilers & Marx, 1996). For each smoothed predictor, four knots were placed at quantiles of the training data; with degree three, omission of the constant (bias) basis function and linear extrapolation beyond the training range, this gives $K = 5$ basis columns per term. No interactions were fitted. The basis dimension was kept small because only 147 respondents had the unsuccessful outcome.

The coefficients γ and θ_{jk} were estimated with the same ridge penalty as in equation (3), applied to standardized basis columns. No derivative-based roughness penalty was used, and no term-specific smoothing parameter was estimated. The models therefore belong to the generalized additive model family (Hastie & Tibshirani, 1986) but are regression-spline models with a common coefficient penalty; they should not be equated with P-spline models (Eilers & Marx, 1996) or with smoothing-parameter selection by restricted maximum likelihood in the mgcv framework (Wood, 2011, 2017).

2.4.4 Model specifications

Five specifications were compared (Table 2). M1 contained the background block alone. M2 added active VLE days and the three transformed click counts as linear terms. M3 used the same information as M2 but replaced active days and the transformed click counts with spline terms. M4 added assessment attempts and average assessment score to M2 as linear terms. M5 used the same information as M4, with spline terms for active days, the transformed click counts and average assessment score and a linear term for assessment attempts. The contrasts M2 – M1 and M4 – M2 therefore measure the incremental value of the engagement and assessment blocks, whereas M3 – M2 and M5 – M4 isolate the change in functional form. All models used the same transformations, outer folds and tuning criterion.

Table 2. Specifications of the fitted prediction models

Model	Predictor blocks	Functional form	Coefficients ^a
M1	Background	Ridge-penalized logistic	4 + dE
M2	Background + engagement	Ridge-penalized logistic	8 + dE
M3	Background + engagement	Additive spline logistic	24 + dE
M4	Background + engagement + assessment	Ridge-penalized logistic	10 + dE
M5	Background + engagement + assessment	Additive spline logistic	30 + dE

Note. Number of penalized coefficients, excluding the intercept; dE is the number of indicator columns for highest education (number of categories minus one). Each spline term contributes five B-spline basis columns. In M3, active VLE days and the three transformed click counts are smoothed; in M5, these four variables and average assessment score are smoothed and assessment attempts enter linearly.

2.5 Repeated nested cross-validation

Predictive performance was estimated by repeated nested cross-validation (Cawley & Talbot, 2010; Stone, 1974; Varma & Simon, 2006). The outer loop used stratified five-fold cross-validation repeated three times with random seeds 2400, 2401 and 2402. Within each outer training set, stratified three-fold cross-validation selected the value of C with the lowest mean Brier score; inner seeds were $700 + 5r + k$, where r and k are the zero-based repeat and outer-fold indices. Category encoding, standardization and knot placement were estimated within each training set only, and the same outer folds were used for all five specifications.

For each repetition, the five held-out prediction vectors were reassembled in the original row order, and each performance measure was calculated once on the complete vector of 1,000 out-of-fold predictions. The reported results are arithmetic means of the three repetition-level values; the 3,000 out-of-fold predictions were not treated as independent observations. Because training sets overlap, no universal unbiased estimator of the variance of a k-fold cross-validation estimate exists (Bengio & Grandvalet, 2004), and cross-validation estimates from small samples can have wide error bars (Varoquaux, 2018). The ranges across repetitions are therefore reported descriptively and are not interpreted as confidence intervals.

2.6 Performance measures

2.6.1 Discrimination

Discrimination was measured by the AUC, which summarizes the trade-off between sensitivity and specificity across all thresholds (Fawcett, 2006); limitations of the AUC as a universal measure of classifier performance are discussed by Hand (2009). Because the unsuccessful outcome is uncommon, average precision (AP) was also reported (Saito & Rehmsmeier, 2015). AP was calculated as the recall-increment-weighted mean of precision,

$$AP = \sum_k (R_k - R_{k-1}) P_k \quad (5)$$

where P_k and R_k are the precision and recall at the kth distinct threshold, rather than as the trapezoidal area under the precision–recall curve. The relationship between receiver operating characteristic (ROC) and precision–recall curves depends on prevalence and on the score distributions (Davis & Goadrich, 2006); a non-informative model has an AP approximately equal to the prevalence (0.147). Differences in AUC are reported descriptively; the nonparametric test for correlated ROC curves (DeLong et al., 1988) was not used for confirmatory testing.

2.6.2 Probabilistic accuracy

$$BS = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{p}_i)^2 \quad (6)$$

where $\hat{p}_i$ is the out-of-fold predicted probability for respondent i. The Brier score is a proper scoring rule (Brier, 1950; Gneiting & Raftery, 2007) and can be decomposed into uncertainty, reliability and resolution components (Murphy, 1973). As a reference, a constant prediction equal to the sample prevalence $\hat{\pi} = 0.147$ gives $BS_{ref} = \hat{\pi}(1 - \hat{\pi}) = 0.1254$. The scaled Brier score,

$$BS_{scaled} = 1 - \frac{BS}{BS_{ref}}, \quad BS_{ref} = \hat{\pi}(1 - \hat{\pi}) \quad (7)$$

expresses the proportional improvement over this constant prediction (Steyerberg et al., 2010). Log loss was also reported:

$$LL = -\frac{1}{n} \sum_{i=1}^n \{y_i \log \hat{p}_i + (1 - y_i) \log (1 - \hat{p}_i)\} \quad (8)$$

for which the constant prediction gives 0.4175. Lower values of BS and LL indicate better probabilistic accuracy.

2.6.3 Calibration

Calibration was assessed with the logistic recalibration model (Miller et al., 1991; Steyerberg et al., 2010)

$$\text{logit } P(Y_i = 1 \mid \hat{p}_i) = a + b \text{ logit } (\hat{p}_i) \quad (9)$$

fitted to the out-of-fold predictions of each repetition, with the intercept a and slope b estimated jointly and averaged across repetitions. Perfect calibration corresponds to a = 0 and b = 1; a slope below 1 indicates predicted risks that are too extreme, and a slope above 1 predicted risks that are too moderate (Van Calster et al., 2019). Because a was estimated jointly with b, it is not the calibration-in-the-large intercept obtained with the slope fixed at 1 (Van Calster et al., 2016). Predictions were clipped to [10⁻⁸, 1 – 10⁻⁸] before the logit transformation. For a visual assessment, the out-of-fold predictions from the first repetition were divided into five approximately equal-sized groups, and the mean predicted risk in each group was plotted against the observed proportion. Because grouped plots can conceal local miscalibration, they were read together with the numerical intercept and slope. Integrated calibration measures (Austin & Steyerberg, 2019) were not computed.

2.7 Paired bootstrap comparisons

For selected contrasts, paired differences in AUC and Brier score were summarized by percentile bootstrap intervals (Efron, 1979) based on 1,000 resamples of respondents from the out-of-fold predictions of the first repetition (seed 910). Both models in each contrast were evaluated on the same resampled indices, preserving the pairing of predictions for each respondent. Because the models were not refitted and the tuning was not repeated within resamples, the intervals are conditional on the prediction vectors of the first repetition and are interpreted descriptively; resampling approaches

that refit the models, such as the .632+ bootstrap (Efron & Tibshirani, 1997), would also reflect variability due to model fitting. The point estimates refer to the first repetition and therefore differ slightly from the differences between the three-repetition means reported in Section 3.2.

2.8 Partial dependence

To describe the fitted relationships, M4 and M5 were each tuned on the complete dataset by five-fold cross-validation (seed 999) and fitted to all 1,000 respondents. For each displayed predictor, every respondent was assigned a common grid value and the predicted probabilities were averaged over the observed values of the remaining predictors; the grids spanned the 5th to 95th percentiles. These partial-dependence curves describe the fitted prediction functions; they are not out-of-sample estimates, causal effects or formal tests of nonlinearity. In particular, a curve on the probability scale can bend even when the corresponding log-odds term is linear, because of the inverse-logit transformation and, for click counts, the transformation applied before modelling.

2.9 Software and reproducibility

All analyses were performed in R. (R-4.6.1) All random seeds are reported above. The outcome definition, predictor blocks, transformations, validation scheme and performance measures were specified before the model comparison was carried out, and reporting was guided by the TRIPOD statement for prediction-model studies (Collins et al., 2015).

3 Results

3.1 Sample characteristics and data quality

All 1,000 responses were complete, and no respondent identifier was duplicated. Pass and Distinction together accounted for 853 respondents (85.3%) and Fail and Withdraw for 147 (14.7%). The largest age band was 0–35 years (632 respondents, 63.2%), followed by 35–55 years (268, 26.8%) and 55 years or above (100, 10.0%). The most common highest-education category was “HE Qualification” (349 respondents, 34.9%). Five records had a sum of click-category counts greater than total clicks; because the questionnaire did not specify whether click categories overlap, these records were retained and flagged.

Table 3 reports medians and interquartile ranges of the numerical variables by outcome group. Compared with successful respondents, unsuccessful respondents had a lower median average assessment score (62.0 vs 72.7), fewer assessment attempts (3 vs 4), fewer active VLE days (43 vs 45) and fewer total VLE clicks (285 vs 330). Median quiz and forum clicks were similar in the two groups, and studied credits and previous attempts had identical medians and interquartile ranges. These comparisons are descriptive.

Table 3. Numerical variables by outcome group: median (interquartile range)

Variable	All (n = 1,000)	Successful(n= 853)	Unsuccessful(n=147)
Studied credits	60.0 (60.0–120.0)	60.0 (60.0–120.0)	60.0 (60.0–120.0)
Previous attempts	0.0 (0.0–0.0)	0.0 (0.0–0.0)	0.0 (0.0–0.0)
Active VLE days	45.0 (40.0–49.0)	45.0 (41.0–49.0)	43.0 (39.0–47.0)
Total VLE clicks	323.5 (191.0–544.2)	330.0 (196.0–571.0)	285.0 (165.0–425.5)
Resource clicks	104.0 (62.8–184.2)	107.0 (64.0–193.0)	88.0 (52.5–154.0)
Quiz clicks	18.0 (15.0–21.0)	18.0 (15.0–21.0)	18.0 (15.5–20.0)
Forum clicks	7.0 (5.0–9.0)	7.0 (5.0–9.0)	6.0 (5.0–9.0)
Assessment attempts	4.0 (3.0–5.0)	4.0 (3.0–6.0)	3.0 (2.0–4.0)
Average assessment score	70.8 (61.7–79.8)	72.7 (63.4–80.6)	62.0 (55.0–71.5)

Note. Successful: Pass or Distinction; unsuccessful: Fail or Withdraw. VLE, virtual learning environment. Resource clicks are reported descriptively and were not used as a predictor.

3.2 Incremental value of predictor blocks

Table 4 summarizes the cross-validated performance of the five models. The background-only model, M1, had an AUC

of 0.543 and a Brier score of 0.1254, indistinguishable from the constant-risk reference (scaled Brier score 0.000); its log loss (0.4177) was also no lower than that of the constant prediction (0.4175). Adding the engagement block (M2) increased the AUC to 0.594 and reduced the Brier score to 0.1240, a 1.1% improvement over the constant prediction. Adding the assessment block produced the largest gain: M4 reached an AUC of 0.736, an AP of 0.318 (about 2.2 times the prevalence), a Brier score of 0.1144 (scaled Brier score 0.088) and a log loss of 0.3732. M4 had the highest AUC and AP and the lowest Brier score and log loss of the five models.

Table 4. Cross-validated performance and calibration of the five models

Model	AUC	AP	Brier	Scaled Brier ^a	Log loss	a	b
M1	0.543	0.176	0.1254	0.000	0.4177	−0.848	0.512
M2	0.594	0.201	0.1240	0.011	0.4110	−0.454	0.732
M3	0.595	0.201	0.1237	0.013	0.4103	−0.357	0.792
M4	0.736	0.318	0.1144	0.088	0.3732	−0.088	0.942
M5	0.721	0.311	0.1154	0.080	0.3768	−0.158	0.904
Constant ^b	0.500	0.147	0.1254	0	0.4175	—	—

Note. Values are means over three repetitions of stratified five-fold outer cross-validation. AUC, area under the receiver operating characteristic curve; AP, average precision; a and b, calibration intercept and slope estimated jointly (ideal values 0 and 1). ^a1 − Brier/0.1254, computed from the rounded mean Brier scores. ^bConstant prediction equal to the sample prevalence (0.147), shown for reference.

3.3 Logistic versus additive spline specifications

Holding the information set fixed, the spline specifications did not improve on their logistic counterparts. With background and engagement information, M3 performed almost identically to M2 (AUC 0.595 vs 0.594; Brier score 0.1237 vs 0.1240). With assessment information added, M5 had a lower AUC (0.721 vs 0.736) and AP (0.311 vs 0.318) and a higher Brier score (0.1154 vs 0.1144) and log loss (0.3768 vs 0.3732) than M4.

3.4 Calibration

All calibration slopes were below 1, indicating predicted risks that were more extreme than the observed outcomes supported. The departure from ideal calibration was largest for M1 (a = −0.848, b = 0.512) and M2 (a = −0.454, b = 0.732) and smallest for M4 (a = −0.088, b = 0.942), followed by M5 (a = −0.158, b = 0.904). The assessment-inclusive logistic model was therefore both the best-discriminating and the best-calibrated specification.

3.5 Paired bootstrap comparisons

Table 5 reports paired differences for selected contrasts in the first repetition. Adding engagement increased the AUC (M2 − M1: 0.047; 95% interval 0.003 to 0.090), whereas the interval for the change in Brier score included zero (−0.0016; −0.0037 to 0.0006). Both intervals for M3 − M2 included zero. Adding assessment information improved both measures (M4 − M2: AUC 0.139, 0.086 to 0.192; Brier score −0.0092, −0.0141 to −0.0041). M5 had a slightly lower AUC than M4 (−0.023; −0.046 to −0.001), with no clear difference in Brier score (0.0014; −0.0012 to 0.0040). Because the models were not refitted within resamples, these intervals are conditional on the first repetition's predictions and are interpreted descriptively.

Table 5. Paired differences in performance with conditional 95% bootstrap intervals

Contrast	ΔAUC (95% interval)	ΔBrier (95% interval)
M2 − M1	0.047 (0.003, 0.090)	−0.0016 (−0.0037, 0.0006)
M3 − M2	−0.007 (−0.040, 0.028)	0.0002 (−0.0020, 0.0024)
M4 − M2	0.139 (0.086, 0.192)	−0.0092 (−0.0141, −0.0041)

Contrast	ΔAUC (95% interval)	ΔBrier (95% interval)
M5 – M4	–0.023 (–0.046, –0.001)	0.0014 (–0.0012, 0.0040)

Note. Differences are the first-named model minus the second, computed on the out-of-fold predictions of the first repetition; a positive ΔAUC or a negative ΔBrier favours the first-named model. Percentile intervals are based on 1,000 paired bootstrap resamples of respondents (seed 910) and are conditional on the prediction vectors (models not refitted). Point estimates therefore differ slightly from the differences between the means in Table 4.

3.6 Graphical evaluation

Figure 1 shows the ROC curves and grouped calibration plots for the five models. The ROC curves of the assessment-inclusive models (M4 and M5) lie clearly above those of M1–M3, consistent with Table 4. In the grouped calibration plot, observed proportions were close to the mean predicted risks in the lower-risk groups for all models. In the highest-risk group, the observed proportions for M4 and M5 were below the mean predicted risk, consistent with calibration slopes below 1. With only five groups per model, the plot cannot reveal local departures from calibration in detail.

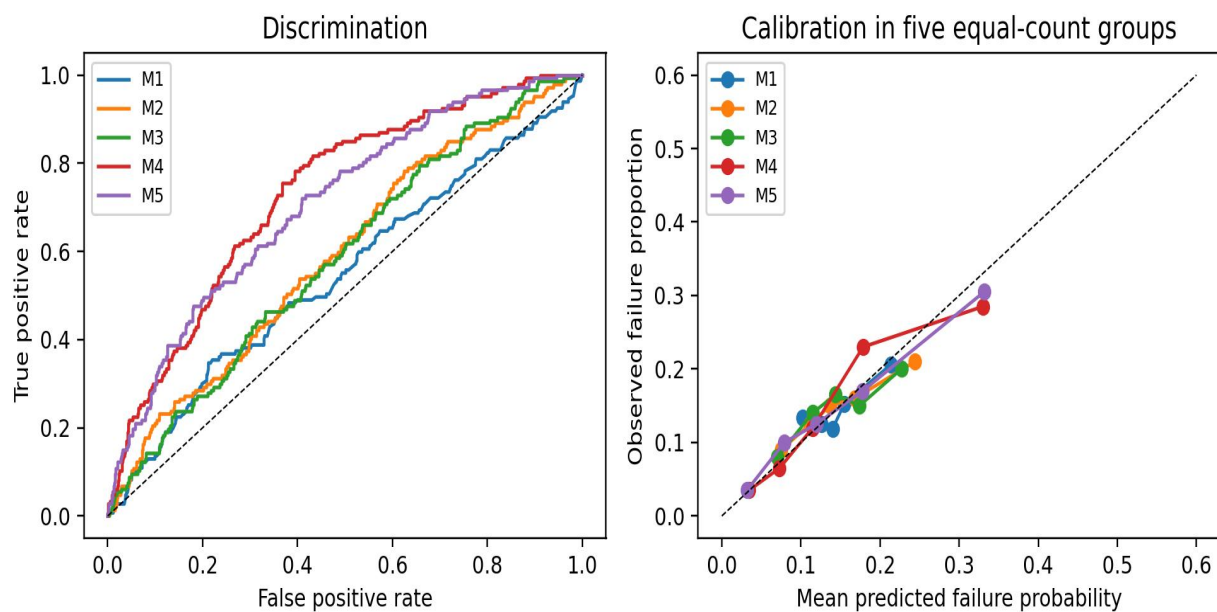


Figure 1. Receiver operating characteristic curves (left) and grouped calibration plots (right) for models M1–M5. Calibration groups are five approximately equal-sized groups of out-of-fold predictions from the first repetition. The dashed line represents a non-informative model (left) or perfect calibration (right). In the axis labels, “failure” denotes the composite unsuccessful outcome (failure or withdrawal).

3.7 Fitted relationships

Figure 2 shows partial-dependence curves for active VLE days, total VLE clicks and average assessment score under M4 and M5. In both models, mean predicted risk decreased as each of these variables increased over most of the displayed range. The curves for average assessment score were similar in the two models, falling from about 0.26–0.27 at a score of 50 to below 0.05 at a score of 95. For active days, the M4 curve declined steadily, whereas the M5 curve was nearly flat, with a slight rise, below about 41 days and declined more steeply thereafter. For total clicks, the M5 curve declined more steeply than the M4 curve above about 600 clicks. Because M5 did not improve validated performance, these differences in shape should not be interpreted as evidence of genuine nonlinearity; they may also reflect correlations among predictors and the chosen basis and penalty.

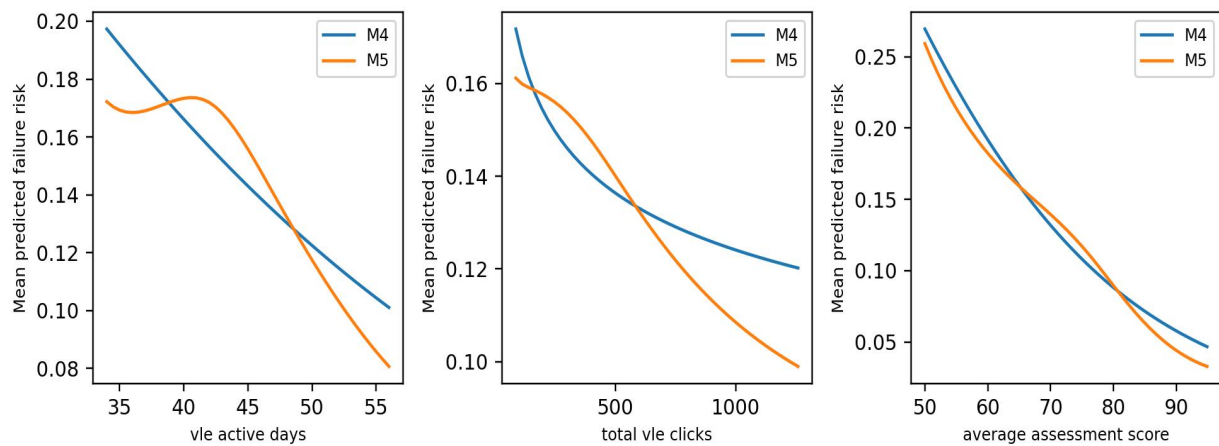


Figure 2. Partial-dependence curves for active VLE days, total VLE clicks and average assessment score under M4 (ridge-penalized logistic) and M5 (additive spline), over the 5th to 95th percentiles of each predictor. Curves show the mean predicted risk of the unsuccessful outcome after tuning and fitting on the complete dataset; they describe the fitted models and are not out-of-sample estimates or causal effects. Confidence bands are not shown.

3.8 Stability across repetitions

Across the three repetitions, the AUC of M4 ranged from 0.731 to 0.743 and its Brier score from 0.1132 to 0.1153; the corresponding ranges for M5 were 0.708 to 0.730 and 0.1142 to 0.1163. The AUC ranges of the two models did not overlap, whereas their Brier-score ranges did. These ranges reflect sensitivity to fold assignment rather than sampling uncertainty.

4 Discussion

4.1 Principal findings

This study compared ridge-penalized logistic and additive spline models for predicting failure or withdrawal in a survey of 1,000 students from Thiruvananthapuram district, using a block-wise design and repeated nested cross-validation. Three findings stand out. First, assessment information carried most of the predictive signal: adding assessment attempts and average score raised the AUC from 0.594 to 0.736 and reduced the Brier score by 0.0096, whereas background variables alone offered no improvement over a constant prediction and engagement variables improved prediction only slightly. Second, spline flexibility did not improve prediction once the information set was fixed: M3 matched M2, and M5 was slightly worse than M4 on every measure. Third, measures beyond the AUC were informative. M1 had an AUC above 0.5 but a Brier score and log loss no better than those of a constant prediction, and the calibration slopes showed that the weaker models produced overly extreme risks.

Even the best model achieved modest absolute performance. Its Brier score was 8.8% lower than that of a constant prediction, and its AP was about twice the prevalence. Whether this level of performance is useful depends on the intended action and on its costs and benefits, which were not assessed here.

4.2 Interpretation in relation to previous research

The dominance of assessment information agrees with reviews that identify assessment grades as common predictors of academic performance (Namoun & Alshanqiti, 2021) and with the finding of Conijn et al. (2017) that LMS data added little once in-between assessment grades were available. The modest contribution of engagement counts is consistent with You's (2016) observation that indicators of meaningful learning behaviour can be more informative than simple frequency counts. In the present data, the engagement variables were aggregate counts whose underlying event definitions were unavailable, so similar numbers may summarize quite different study behaviours (Gašević et al., 2015). Because predictive relationships vary across courses and instructional conditions (Conijn et al., 2017; Gašević et al., 2016), the relative contributions of the predictor blocks found here may not transfer to other institutions or course designs.

That the additive models did not outperform the logistic models is consistent with the limited information available for estimating flexible functions. With 147 unsuccessful outcomes, M5 estimated 30 + dE penalized coefficients, compared with 10 + dE for M4 (Table 2). Approximating the anticipated Cox–Snell R^2 by the value implied by the cross-validated log loss of M4 (about 0.085), criterion (i) of Riley et al. (2020), which targets an expected shrinkage factor of at least 0.9, implies roughly 101 participants per estimated coefficient, that is, more than 1,000 respondents for M4 and more than 3,000 for M5. The sample therefore falls short of this guide, particularly for the flexible models, which is consistent with the need for penalization and with the absence of a gain from the splines (see also van Smeden et al., 2016, 2019). This calculation is approximate, because it uses a cross-validated rather than an anticipated R^2 and because ridge penalization partly offsets the effect of a small sample. When added flexibility does not improve validated performance, a transparent and parsimonious model is preferable, particularly when predictions inform decisions about individuals (Rudin, 2019). The comparison concerns prediction rather than explanation (Breiman, 2001; Shmueli, 2010): that the logistic form predicts at least as well does not establish that the underlying relationships are linear.

The calibration results support the view that calibration is a weak point of predictive analytics that must be checked

rather than assumed (Van Calster et al., 2019). Slopes below 1, most pronounced for M1 and M2, indicate that the predicted risks of these models varied more than the outcomes justified, even after ridge penalization tuned by the Brier score.

4.3 Implications for statistical practice

Three implications follow. First, separating information blocks from functional form clarifies where predictive gains originate; in this study, virtually all of the improvement came from additional information rather than from flexibility. Second, proper scoring rules and calibration should accompany the AUC when risk models are compared, because a model can rank above chance without improving probabilistic accuracy (Gneiting & Raftery, 2007; Steyerberg et al., 2010). Third, tuning and evaluation should be separated, for example by nested cross-validation, to avoid optimistic selection bias (Cawley & Talbot, 2010; Varma & Simon, 2006), and the uncertainty of cross-validated estimates should be acknowledged rather than ignored (Bengio & Grandvalet, 2004; Varoquaux, 2018).

4.4 Future research

Future work should (i) collect dated measurements so that predictions can be made at defined points before the outcome and validated temporally; (ii) validate the models externally in other districts and types of institution; (iii) compare fully penalized additive models with smoothing parameters estimated by restricted maximum likelihood (Wood, 2011, 2017), as well as lasso and elastic-net penalties; (iv) model failure and withdrawal separately; (v) assess calibration with flexible calibration curves and integrated measures (Austin & Steyerberg, 2019) and evaluate performance in subgroups; and (vi) use simulation studies designed according to established guidance (Burton et al., 2006; Morris et al., 2019) to identify the sample sizes and signal structures under which additive flexibility improves the prediction of uncommon academic outcomes. Before any operational use, a risk score would also require stakeholder review, a defined intervention protocol, human oversight and an evaluation of its consequences for students (Alfredo et al., 2024; Ifenthaler & Yau, 2020; Pardo & Siemens, 2014; Slade & Prinsloo, 2013).

5 Conclusions

In this survey, assessment information provided most of the predictive value for failure or withdrawal, engagement counts and background variables alone performed no better than a constant prediction in probabilistic terms. Within each information set, additive spline models did not improve on ridge-penalized logistic regression: the assessment-inclusive logistic model achieved the best cross-validated performance (AUC 0.736, average precision 0.318, Brier score 0.1144, log loss 0.3732) and the calibration closest to ideal (intercept -0.088 , slope 0.942). In this setting, with a limited number of unsuccessful outcomes, the parsimonious penalized logistic model is preferable to the more flexible alternative, and calibration and proper scoring rules should be reported alongside the AUC.

References

1. Akçapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education*, 16, Article 40. <https://doi.org/10.1186/s41239-019-0172-z>
 2. Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6, Article 100215. <https://doi.org/10.1016/j.caeai.2024.100215>
 3. Alyahyan, E., & Düşteğör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17, Article 3. <https://doi.org/10.1186/s41239-020-0177-7>
 4. 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). ACM. <https://doi.org/10.1145/2330601.2330666>
 5. Austin, P. C., & Steyerberg, E. W. (2019). The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine*, 38(21), 4051–4065. <https://doi.org/10.1002/sim.8281>
 6. Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3–17. <https://doi.org/10.5281/zenodo.3554657>
 7. Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, 5, 1089–1105. <https://www.jmlr.org/papers/v5/grandvalet04a.html>
 8. Bodily, R., & Verbert, K. (2017). Review of research on student-facing learning analytics dashboards and educational recommender systems. *IEEE Transactions on Learning Technologies*, 10(4), 405–418. <https://doi.org/10.1109/TLT.2017.2740172>
 9. Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231. <https://doi.org/10.1214/ss/1009213726>
 10. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
 11. Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24), 4279–4292. <https://doi.org/10.1002/sim.2673>
 12. Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11, 2079–2107.
- Volume-12 | Issue-04 | October 2026

13. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ*, 350, Article g7594. <https://doi.org/10.1136/bmj.g7594>
14. Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*, 10(1), 17–29. <https://doi.org/10.1109/TLT.2016.2616312>
15. Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), 215–232. <https://doi.org/10.1111/j.2517-6161.1958.tb00292.x>
16. Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). ACM. <https://doi.org/10.1145/1143844.1143874>
17. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3), 837–845. <https://doi.org/10.2307/2531595>
18. Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26. <https://doi.org/10.1214/aos/1176344552>
19. Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438), 548–560. <https://doi.org/10.1080/01621459.1997.10474007>
20. Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89–121. <https://doi.org/10.1214/ss/1038425655>
21. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
22. Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), 304–317. <https://doi.org/10.1504/IJTEL.2012.051816>
23. Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), 59–109. <https://doi.org/10.3102/00346543074001059>
24. Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22. <https://doi.org/10.18637/jss.v033.i01>
25. Gašević, D., Dawson, S., Rogers, T., & Gašević, 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. <https://doi.org/10.1016/j.iheduc.2015.10.002>
26. Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64–71. <https://doi.org/10.1007/s11528-014-0822-x>
27. Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. <https://doi.org/10.1198/016214506000001437>
28. Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1), 103–123. <https://doi.org/10.1007/s10994-009-5119-5>
29. Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3), 297–310. <https://doi.org/10.1214/ss/1177013604>
30. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: A systematic literature review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education* (pp. 175–199). ACM. <https://doi.org/10.1145/3293881.3295783>
31. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67. <https://doi.org/10.1080/00401706.1970.10488634>
32. Ifenthaler, D., & Yau, J. Y.-K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development*, 68(4), 1961–1990. <https://doi.org/10.1007/s11423-020-09788-z>
33. Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1), 6–47. <https://doi.org/10.18608/jla.2014.11.3>
34. Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics dataset. *Scientific Data*, 4, Article 170171. <https://doi.org/10.1038/sdata.2017.171>
35. Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an “early warning system” for educators: A proof of concept. *Computers & Education*, 54(2), 588–599. <https://doi.org/10.1016/j.compedu.2009.09.008>
36. Miller, M. E., Hui, S. L., & Tierney, W. M. (1991). Validation techniques for logistic regression models. *Statistics in Medicine*, 10(8), 1213–1226. <https://doi.org/10.1002/sim.4780100805>
37. Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102. <https://doi.org/10.1002/sim.8086>
38. Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4), 595–600. [https://doi.org/10.1175/1520-0450\(1973\)012<0595:ANVPOT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2)
39. Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1), Article 237. <https://doi.org/10.3390/app11010237>
40. Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society:*

41. Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3), 438–450. <https://doi.org/10.1111/bjet.12152>
42. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85), 2825–2830. <https://www.jmlr.org/papers/v12/pedregosa11a.html>
43. Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, 16(4), 385–407. <https://doi.org/10.1007/s10648-004-0006-x>
44. Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Jr., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G. S., & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, 368, Article m441. <https://doi.org/10.1136/bmj.m441>
45. Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1), 135–146. <https://doi.org/10.1016/j.eswa.2006.04.005>
46. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), Article e1355. <https://doi.org/10.1002/widm.1355>
47. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
48. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), Article e0118432. <https://doi.org/10.1371/journal.pone.0118432>
49. Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. <https://doi.org/10.1214/10-STS330>
50. Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10), 1380–1400. <https://doi.org/10.1177/0002764213498851>
51. Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 252–254). ACM. <https://doi.org/10.1145/2330601.2330661>
52. Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10), 1510–1529. <https://doi.org/10.1177/0002764213479366>
53. Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerd, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, 21(1), 128–138. <https://doi.org/10.1097/EDE.0b013e3181c30fb2>
54. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 111–133. <https://doi.org/10.1111/j.2517-6161.1974.tb00994.x>
55. 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. <https://doi.org/10.1016/j.chb.2014.05.038>
56. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
57. 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>
58. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., & Topic Group “Evaluating diagnostic tests and prediction models” of the STRATOS initiative. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17, Article 230. <https://doi.org/10.1186/s12916-019-1466-7>
59. Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., & Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: From utopia to empirical data. *Journal of Clinical Epidemiology*, 74, 167–176. <https://doi.org/10.1016/j.jclinepi.2015.12.005>
60. van Smeden, M., de Groot, J. A. H., Moons, K. G. M., Collins, G. S., Altman, D. G., Eijkemans, M. J. C., & Reitsma, J. B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, 16, Article 163. <https://doi.org/10.1186/s12874-016-0267-3>
61. van Smeden, M., Moons, K. G. M., de Groot, J. A. H., Collins, G. S., Altman, D. G., Eijkemans, M. J. C., & Reitsma, J. B. (2019). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, 28(8), 2455–2474. <https://doi.org/10.1177/0962280218784726>
62. Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, Article 91. <https://doi.org/10.1186/1471-2105-7-91>
63. Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180, 68–77. <https://doi.org/10.1016/j.neuroimage.2017.06.061>
64. Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98–110. <https://doi.org/10.1016/j.chb.2018.07.027>
65. Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, 2(2), 5–13. <https://doi.org/10.18608/jla.2015.22.2>
66. Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1), 3–36. <https://doi.org/10.1111/j.1467-9868.2010.00749.x>

67. Wood, S. N. (2017). Generalized additive models: An introduction with R (2nd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9781315370279>
68. You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education*, 29, 23–30. <https://doi.org/10.1016/j.iheduc.2015.11.003>
69. Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2
70. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>

EPH – International Journal of Mathematics and Statistics

<https://ejjms.com/index.php/ms/index>



This work is licensed under a Creative Commons Attribution 4.0 International License. The authors retain ownership of the copyright for their article, You are free to: Share — copy and redistribute the material in any medium or format for any purpose, even commercially. To see the complete license contents,

please visit <http://creativecommons.org/licenses/by/4.0/>.