Future Technology
Open Access Journal
ISSN 2832-0379

February 2027| Volume 06 | Issue 01 | Pages 35-50

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.6.1.4

Future Publishing LLC

Article

Explainable machine learning for early identification of weak professional identity among pre-service early childhood teachers in China

Meijing Chen[1], Kah Choon Low[2]*, Zarina Zulkifli[3], Say Lee Chong[4], Peizhi Zheng[5,6]

[1]Ghazali Shafie Graduate School of Government, Universiti Utara Malaysia, Kedah, Malaysia
[2]School of Government (SoG) UUM COLGIS and Institute of Local Government Studies (iLGS), Universiti Utara Malaysia, 06010 UUM Sintok Kedah, Malaysia
[3]School of Government (SoG), UUM COLGIS, Universiti Utara Malaysia, Kedah, Malaysia
[4]Professional and Continuing Education Centre (PACE), Universiti Utara Malaysia and Independent Researcher cum e-Commerce Entrepreneur, Kedah, Malaysia
[5]Xiamen S & M Imp. & Exp. Co., Ltd., Xiamen, Fujian, China
[6]International One Belt One Road Infrastructure Chamber of Commerce, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 18 July 2026
Received in revised form
29 October 2026
Accepted 18 November 2026

Keywords:
Explainable artificial intelligence, XGBoost, SHAP, Professional identity, Pre-service teachers, Early warning

*Corresponding author
Email address:
kahchoon@gw.uum.edu.my

DOI: 10.55670/fpll.futech.6.1.4

ABSTRACT

Teacher education programs rarely know which of their students will finish training with a weak sense of themselves as teachers, although such students may be more likely to leave the profession early. This study examined whether weak professional identity among pre-service early childhood teachers in China can be predicted one or two survey waves in advance. We analyzed an open three-wave panel of 1,646 students, trained logistic regression, random forest, and XGBoost models on measures of teacher support, teaching efficacy, and identity, and interpreted the predictions with SHapley Additive exPlanations (SHAP). XGBoost reached an area under the curve (AUC) of 0.712 from first-wave data and 0.749 from second-wave data, and its predicted risks were well calibrated. Logistic regression performed almost as well. Prior identity accounted for most of the predictive accuracy, since a model using first-wave identity alone reached 0.709. Support and efficacy sharpened the separation of individual risk estimates and showed where each student's risk came from. A model built from five subscales of support and efficacy plus demographics, without any identity items, still reached 0.677. Volitional identity, a student's commitment to choosing teaching, was the strongest predictor, and students scoring below 3.0 on it were nearly three times as likely as others to report weak identity two waves later. A program able to counsel one student in five would reach about two in five of the students later at risk, and about half of the students it selected would be true cases.

1. Introduction

The rapid growth of early childhood education in China has turned attention to how kindergarten teachers are trained, since their preparation shapes the quality of what children receive [1]. Teachers' understanding of their own role shapes how they work with young children and families [2]. Much of that understanding takes form during initial teacher education, before the first job [3,4]. In a study of pre-service and beginning teachers, Hong [5] found that those who had already left teaching reported the most emotional burnout. Burnout and low job satisfaction are also associated with stronger intentions to quit among practicing teachers [6]. Students who graduate with little commitment to teaching may feel these pressures more sharply in their first years. Whether a particular identity score at graduation predicts leaving, however, has yet to be shown in data that follow students into employment. For a teacher education program, the difficulty is mostly one of timing, since identity, when it is measured at all, tends to be surveyed now and then and reported as group averages. By the time results are available, the students concerned have often moved on. Tutors need an earlier indication of which students are drifting towards a weak identity and what lies behind it. Prediction models trained on routine questionnaire data can

provide such an indication, and SHapley Additive exPlanations (SHAP) can then show which answers raised or lowered each student's predicted risk [7,8]. Research on pre-service teachers' professional identity has mostly estimated average relationships in a single sample, linking identity to learning climate, self-efficacy, motivation, engagement, emotional labor and perceived support [9-12]. Such averages describe the typical student and leave a program unable to say which of its own students need attention this term. Machine learning has rarely been applied to pre-service teachers. In the one example we found, Ann and Boonnam [13] compared seven classifiers for predicting academic performance, not identity, and did not report calibration or the practical value of the predictions. The panel used in the present study was collected for a cross-lagged panel network analysis of teaching efficacy, teacher support, and identity [14]. This design describes how the three constructs relate to one another over time across the whole group, but it does not estimate risk for individual students.

We set three objectives for the study: first, how accurately weak identity at a later wave can be predicted from support, efficacy, and identity measured one or two waves earlier, and how much support and efficacy add once prior identity is known. We also wanted to identify the predictors that carry most weight and to test whether thresholds or interactions in their effects add anything beyond a linear model. Finally, we asked what a screen built on these predictions would be worth to a program that can counsel only some of its students. Because the panel is openly licensed, a program can try the approach before collecting its own data.

2. Analysis of the problem and related work

2.1 Professional identity of pre-service early childhood teachers

Professional identity refers to how students see themselves as teachers, including the value they attach to the profession and their willingness to commit to it [3,15]. Reviews describe this identity as changing during training, shaped by coursework, placements, and relationships with teacher educators [16,17]. Questionnaire instruments distinguish several domains, including motivation, commitment, and self-efficacy [18]. Self-efficacy, motivation, and commitment were closely related when Canrinus et al. [19] examined them together, supporting a model that treats efficacy and identity jointly. Pre-service early childhood teachers in China face particular pressures, and in a four-year study, students more aware of the stigma attached to their future occupation reported a weaker occupational identity later on [20]. Emotional labor adds a further strain, since its link with weaker identity in this group runs through emotional exhaustion [11]. Government-funded students form a special group, because their places come with financial support and a duty to teach after graduation. Among these students, those less satisfied with the policy and less supported by their teachers fell into the profiles with the lowest identity [12].

2.2 Teacher support and teaching efficacy as resources for identity

Support from teacher educators is one of the two resources most often linked to identity. Self-determination theory explains this link, holding that students who feel cared for by their teachers are more likely to adopt an occupation's values as their own [21]. Studies of pre-service teachers agree with this view, both regarding learning climate in interactions with professors [9] and perceived teacher support [12]. Support can take several forms, from help with learning to help in building ability to emotional reassurance, and these forms need not have the same effect. Bandura [22] argued that people persist with tasks they expect to master, and teaching efficacy carries that idea into the classroom. Gibson and Dembo [23] applied this idea to teaching and separated two beliefs. Personal teaching efficacy is a teacher's confidence in her own ability to teach, while general teaching efficacy is the belief that teaching can overcome what children bring from home. The same two factors appeared among Chinese teachers [24]. Efficacy has been tied to occupational commitment and to a lower intention to quit in both pre-service and practicing teachers [25,26], and it has been studied as a source of identity among pre-service early childhood teachers [10]. Almost all of this evidence comes from linear models fitted at a single time point. Such a model assumes that one more point of support or efficacy matters equally at every level of the scale. Teacher educators often need to know something else. They want to know the score below which a student needs attention, and whether weak support matters more for a student who also doubts her teaching. Flexible models can capture thresholds and interactions of this kind. Before any threshold is taken at face value, however, the results should be compared with linear models that also allow curvature and interaction.

2.3 Machine learning and explainability in educational early warning

Early-warning systems are best known from research on school dropout, where a review by Bowers et al. [27] found that many published flags identified students with low precision, limiting their use when only a few students can be helped. Educational data mining has since turned to machine learning [28]. For questionnaire data, gradient boosted trees [29,30] remain among the most accurate methods [31]. TreeSHAP [8] computes exact Shapley values for these models, from which one fitted model yields both a ranking of predictors and an explanation for each student. Studies have used this approach to predict college students' reliance on generative artificial intelligence [32] and to find threshold-like effects in students' intention to use e-learning [33]. Christodoulou et al. [34] found no general advantage of machine learning over logistic regression across clinical prediction models, and comparisons that tune the flexible model on the test data overstate its accuracy [35]. A model that decides who receives counseling must also give risks that match observed rates and must do better than referring every student or none [36,37]. Explanations that shift from one analysis to the next, or that tutors cannot follow, are unlikely to be trusted [38,39]. Each of these points is now part of current reporting guidance for prediction studies [40], and work on algorithmic bias adds that errors should be compared across groups of students [41]. Where a logistic coefficient gives one slope for each predictor, the same for every student, a SHAP value describes one student and depends on her other answers as well as on the predictor in question [8,42]. Rudin [39] warned that explanations added after fitting can mislead when a simpler interpretable model would predict as well. For this reason, we kept logistic regression as a full benchmark throughout and limited the boosted trees to a depth of two. Shallow trees keep the explanations close to how the model actually combines its inputs.

2.4 Gaps in existing work and problem statement

Table 1 compares the present study with the closest earlier work. The identified studies used cross-sectional data or small panels, applied linear models, and reported associations for the group as a whole. Of the two machine learning studies, one predicted academic performance without explaining its forecasts [13], and the other used SHAP to find threshold-like effects in the intention to use e-learning [33]. Neither addressed teacher identity or decisions to refer. We found no study that combined a longitudinal design, out-of-sample prediction, explanation, calibration, and decision analysis for this outcome. The cross-lagged panel network analysis of the present data [14] comes closest in design, but it describes the group as a whole, whereas screening needs a risk estimate for each student. Teacher education programs in China therefore lack a transparent screening tool to identify, early in training, students likely to graduate with a weak professional identity. This study builds such a tool from open panel data and validates it internally. We compare it with simpler models, including one that uses prior identity alone, and check how well its risks are calibrated and how useful it is for referral decisions. We also examine what drives its predictions for the group and for individual students, whether it works equally well across subgroups, and whether the main results hold in an independent analysis in R.

3. Materials and methods

Figure 1 summarizes the study workflow. All steps used the deposited data file unchanged, and the code for every number and figure is provided with the manuscript.

3.1 Data source and participants

The data come from an open three-wave panel of Chinese pre-service early childhood teachers, deposited in Zenodo under a CC BY 4.0 license [14]. We downloaded the file on 1 October 2026 and confirmed that its MD5 checksum (ea65634febe758427fbc7ebfbf99ff3e) matched the published value. The file contains 1,646 respondents with scale scores for perceived teacher support, teaching efficacy, and professional identity at each wave, along with gender, only-child status, and place of origin. Scores coded as -999 were treated as missing. Of the 1,646 respondents, 674 took part in all three waves, 425 in the first two only, 346 in the last two only, and 201 in the first and third only (Table 2). The deposit does not record the dates of the waves or the time between them. The practical lead time of the two horizons is therefore unknown, and a program should check the spacing of its own surveys before treating a first-wave forecast as a warning given a semester or more in advance. Grade and age were not used as predictors.

Table 1. Comparison of the present study with closely related studies of pre-service teacher identity and explainable prediction in education

Study	Sample and setting	Design	Outcome	Analytic approach
Olsen et al. [9]	190 early childhood teacher students	Cross-sectional survey	Professional identity	Regression on learning climate and academic self-efficacy
Liu and Xin [11]	654 pre-service early childhood teachers	Cross-sectional survey	Professional identity	Moderated mediation
Tian et al. [12]	620 government-funded pre-service preschool teachers, China	Cross-sectional survey	Professional identity	Latent profile analysis and regression
Ren et al. [20]	34 pre-service early childhood education teachers, China	Four-wave panel	Occupational identity	Repeated measures and cross-lagged panel analysis
Ann and Boonnam [13]	Pre-service teachers in five groups	Prediction with fivefold cross-validation	Academic performance	Seven classifiers with SMOTE
Słowiński et al. [33]	University students, Poland	Re-analysis of survey data	Intention to use e-learning	Linear benchmark and explainable machine learning with SHAP
This study	1,646 pre-service early childhood teachers, China	Three-wave panel, two horizons, repeated nested cross-validation	Weak professional identity	Three classifiers, persistence baseline, TreeSHAP, calibration, decision curves, subgroup audit, leave-cohort-out validation, R replication

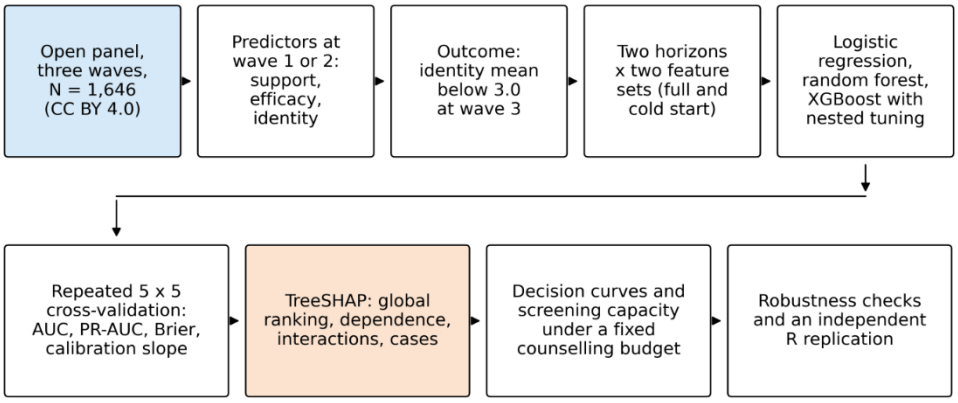


Figure 1. Workflow of the study, from the open three-wave panel and the definition of the outcome, through the prediction tasks and model comparison, to the SHAP explanations, decision analysis, robustness checks and the independent R replication

Table 2. Sample characteristics and first-wave descriptive statistics

Characteristic	Range	n (%) or mean	SD
Respondents in the file		1,646	
Female		1,532 (93.1%)	
Only child		288 (17.5%)	
Place of origin, code 2		1,151 (69.9%)	
Responded at waves 1, 2 and 3		1,300, 1,445, 1,221	
Learning support (wave 1)	1 to 6	4.59	0.65
Ability support (wave 1)	1 to 6	3.91	0.88
Emotional support (wave 1)	1 to 6	4.24	0.64
Personal teaching efficacy (wave 1)	1 to 6	4.06	0.60
General teaching efficacy (wave 1)	1 to 6	4.16	0.71
Intrinsic value identity (wave 1)	1 to 4	3.39	0.49
Extrinsic value identity (wave 1)	1 to 4	3.25	0.49
Volitional identity (wave 1)	1 to 4	3.16	0.56
Professional identity, overall (wave 1)	1 to 4	3.29	0.47
Professional identity, overall (wave 3)	1 to 4	3.20	0.48

Note: Counts refer to all 1,646 respondents in the deposited file. The respondent counts for waves 1, 2 and 3 are those with an overall identity score at that wave. Means and SDs refer to the 875 students in the first-horizon sample. Labels for the two place-of-origin codes were not included in the deposit.

The grade codes ran from 1 to 15 without labels, and age contained implausible values. We used grade only to mark cohorts for the leave-cohort-out validation. In the first-horizon sample, all but two students had code 1 or 2; these two codes formed the held-out cohorts, while the remaining two students were left out of this check. In the second-horizon sample, codes 1 and 2 formed two cohorts, and the other codes were pooled into a third. The deposit has no class or institution codes, leaving cohort as the only available grouping.

3.2 Measures

Perceived teacher support was recorded as three subscale means for learning support, ability support, and emotional support, each on a six-point scale. Teaching efficacy was recorded as personal and general teaching efficacy, also on six-point scales, following the two-factor structure of Yu, Xin, and Shen [24]. Professional identity was recorded as an overall mean and three facet means on a four-point scale. In the deposit, the facets are labeled intrinsic value identity, extrinsic value identity, and volitional identity. Higher scores mean more support, greater efficacy, and stronger identity. Because the file contains only subscale scores, internal consistency coefficients such as Cronbach's alpha could not be computed for this sample. Programs that apply the approach locally should collect item-level responses if possible or at least report the published reliability of the Chinese versions of the instruments before using the scores for screening.

3.3 Outcome definition

The outcome was weak professional identity at the third wave, defined as an overall identity mean below 3.0. On the four-point scale, 3.0 is the point at which a student agrees with the identity statements on average. A lower mean shows that, overall, she disagreed with them. Scores were heaped at exactly 3.0 at every wave, with 28.2%, 36.3% and 32.3% of respondents at that value (Figure 2). Used as a continuous outcome, the score would have been dominated by this heap. The cut-off instead separates a lower tail that held 19.2%, 19.4% and 24.2% of respondents at the three waves. Checking the cut-off against an external criterion would need a measure of intention to leave or of later career choice, which lies outside the deposit. We examined its influence on the results with alternative cut-offs from 2.8 to 3.2, an ordinal outcome, and the continuous score.

3.4 Prediction tasks and feature sets

The prediction tasks cross two forecast horizons with two feature sets (Table 3). In the first horizon, features measured at the first wave were used to predict the outcome two waves later. In the second horizon, second-wave features predicted it one wave later. The full feature set contained the three support subscales, the two efficacy subscales, the three identity facets, and three binary demographic indicators. For the cold start set, we left out the identity facets to represent a program that measures support and efficacy but has never run an identity survey. Each task used every respondent with complete features at the predictor wave and an observed outcome at the third wave.

3.5 Models and tuning

Logistic regression with an L2 penalty (C = 1) on standardized features was the transparent benchmark. A random forest [43] with 500 trees and at least five observations per leaf represented bagging, and XGBoost [29] represented boosting. XGBoost builds an additive ensemble of K regression trees, and the predicted log-odds for student i is:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \mathcal{F} \tag{1}$$

where F is the space of regression trees. Its trees are learned by minimizing the regularized objective.

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \left[\gamma T_k + \frac{\lambda}{2}\|w_k\|^2\right] \tag{2}$$

where l is the logistic loss, and T_k is the number of leaves of tree k. The vector w_k holds the leaf weights of tree k, and γ and λ penalize tree complexity [29]. We fixed the number of trees at 200, row and column subsampling at 0.8, and λ at 1. Maximum depth, learning rate, and minimum child weight were tuned by grid search inside each training fold with three-fold stratified cross-validation, which keeps the tuning away from the test data [35]. Table 4 lists the search space and the values chosen most often.

3.6 Validation and evaluation metrics

Performance was estimated with five-fold stratified cross-validation repeated five times, giving 25 test folds for each model and task. Discrimination was measured by the area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (PR-AUC).

Table 3. Prediction tasks defined by forecast horizon and feature set, with sample size and prevalence of weak identity

Task	Predictor wave	Features	Number	n	Weak identity at wave 3
Wave 1, full	Wave 1	Support (3), efficacy (2), identity (3), demographics (3)	11	875	213 (24.3%)
Wave 2, full	Wave 2	Support (3), efficacy (2), identity (3), demographics (3)	11	1,020	249 (24.4%)
Wave 1, cold start	Wave 1	Support (3), efficacy (2), demographics (3)	8	875	213 (24.3%)
Wave 2, cold start	Wave 2	Support (3), efficacy (2), demographics (3)	8	1,020	249 (24.4%)

Table 4. XGBoost hyperparameter search space and the values selected most often by nested tuning

Hyperparameter	Search values	Modal selection
max_depth	2, 3	2
learning_rate	0.03, 0.10	0.03
min_child_weight	1, 5	1
n_estimators	200 (fixed)	200
subsample, colsample_bytree	0.8 (fixed)	0.8
reg_lambda	1 (fixed)	1

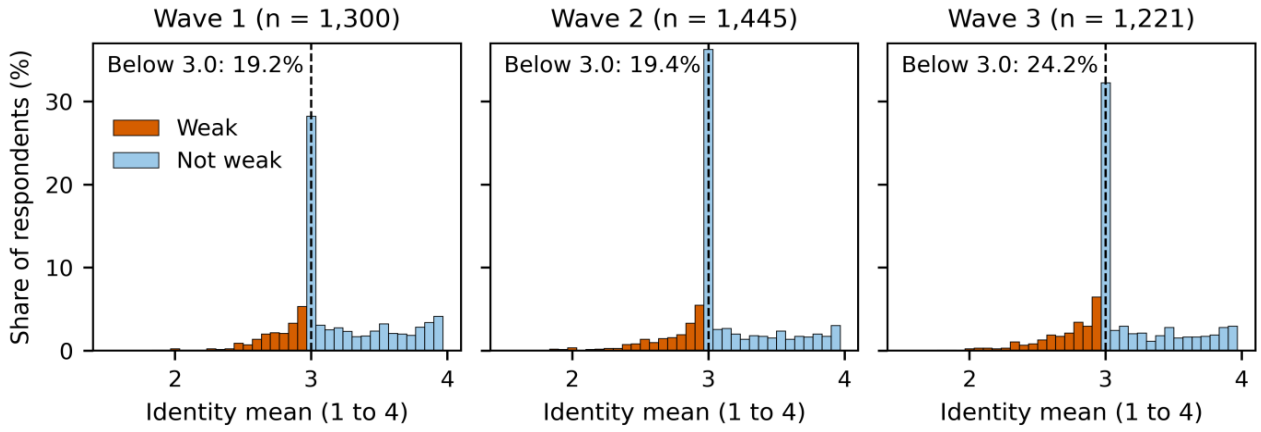


Figure 2. Distribution of the overall professional identity mean at each of the three waves, showing the heap of responses at exactly 3.0. Orange bars mark weak identity (a mean below 3.0), light blue bars mark the remaining students, and the dashed line marks 3.0

PR-AUC is more informative than accuracy when the positive class is a minority [44]. The accuracy of the predicted probabilities was measured by the Brier score.

$$BS = \frac{1}{n}\sum_{i=1}^{n} (p_i - y_i)^2 \quad (3)$$

where p_i is the predicted probability and y_i the observed outcome. Calibration was summarized by the calibration slope, obtained by regressing the outcome on the logit of the pooled out-of-fold predictions, and was also inspected across deciles of predicted risk [36]. Differences in AUC between XGBoost and the other models were tested with the corrected resampled t-test [45]. This test widens the variance to allow for the overlap between training sets. Practical value was assessed with decision curve analysis [37]. For a referral threshold p_t, the net benefit of a model is:

$$NB(p_t) = \frac{TP}{n} - \frac{FP}{n} \times \frac{p_t}{1-p_t} \quad (4)$$

where TP and FP are the numbers of true and false positives among students whose predicted risk is at least p_t. A model is useful at p_t when its net benefit is higher than that of referring every student and that of referring none. Because no resampling of the minority class was applied, the predicted probabilities keep the observed base rate and can be read as risks. To express screening capacity in a form programs can use, we ranked students by predicted risk and flagged the top 10%, 20%, 30%, and 40%. For each budget, we recorded the share of later cases captured and the share of flagged students who were true cases.

3.7 Explainability analysis

For a model f and feature set M, the Shapley value of feature j for student i is:

$$\varphi_j(i) = \sum_{S \subseteq M \setminus \{j\}} \frac{|S|!(|M|-|S|-1)!}{|M|!} \left[f_{S\cup\{j\}}(x_i) - f_S(x_i) \right] \quad (5)$$

Each student's prediction equals a baseline plus the sum of the feature contributions [7]. We applied TreeSHAP [8] to an XGBoost model refitted on the full sample of each task with the model-tuned settings and expressed contributions on the log-odds scale. TreeSHAP was run in its path-dependent (conditional) form, not the interventional form, using SHAP 0.51.0 with XGBoost 3.2.0. This form uses conditional expectations from the fitted trees and requires no background sample. Because the explained model saw all students, its explanations are slightly optimistic relative to the out-of-fold performance, and the bootstrap refits described below are the main check on their stability. When

predictors are correlated, Shapley values divide shared information among them. The attributions of related features, such as the three support subscales, are therefore not independent of one another [46]. Global importance was summarized by the mean absolute SHAP value of each feature. Dependence plots show how each feature's value relates to its contribution, and the value at which a smoothed dependence curve first crosses zero was taken as an empirical threshold. Observed outcome rates were then compared on either side of it. SHAP interaction values measured how much of the explanation came from pairs of features. To test stability, we refit the model on 20 bootstrap samples, compared each importance ranking with the full-sample ranking by Kendall's tau, and recorded where the dependence curves crossed zero in each refit. These crossings give a range for each threshold.

3.8 Robustness checks, subgroup audit and independent replication

Most further checks used the first-horizon full model. To see how much the other predictors add to prior identity, we fitted a persistence baseline, a logistic regression on prior overall identity alone, at both horizons, together with a second baseline on prior weak status. The full models were compared with this baseline by the change in AUC, the category-free net reclassification improvement (NRI), and the integrated discrimination improvement (IDI) [47,48]. These comparisons used fixed modal hyperparameters for XGBoost and bootstrap intervals for NRI and IDI. A further model added only the three support and two efficacy subscales to prior identity, to isolate their contribution. Straightlining, meaning identical scores on all three identity facets, was flagged at the predictor wave [49], and flagged respondents were removed. We also moved the cut-off from 2.8 to 3.2, fitted an ordered logistic regression to a three-level outcome (below 3.0, 3.0 to below 3.5, and 3.5 or above) scored by its out-of-fold concordance, and predicted the continuous third-wave score with linear regression and XGBoost. Whether a linear model misses thresholds or interactions was tested directly. Logistic regression with cubic regression splines, logistic regression with the two strongest pairwise interactions, and explainable boosting machines [50] with and without pairwise interactions were compared with the main models, and likelihood-ratio tests asked whether splines or interactions improved the logistic fit. Inverse probability of retention weights corrected for selective attrition between the first and third waves [51]. First-wave and second-wave features were combined for respondents present at all three waves, and a label permutation test with 200 permutations checked that the observed AUC exceeded chance. Because random folds can place students from the same cohort in both training and test sets, each model was also trained on all but one cohort and tested on the left-out cohort [52]. Finally, the out-of-fold predictions of both full models were audited by gender, only-child status, and place of origin, comparing the AUC, mean predicted risk, observed rate, and share flagged under a 20% budget within each subgroup [41].

We also replicated the three main tasks independently in R. The deposited SPSS file was read with haven, and gradient boosting was fitted with gbm [30], random forests with ranger and logistic regression with base glm, all under five-fold cross-validation repeated five times with new fold assignments. Shapley values for the gbm models were estimated by permutation sampling in base R [42], without any code shared with the Python analysis. The R analysis also covered the persistence baseline, the interaction test, variance inflation factors, the leave-cohort-out validation, the predictor-wave straightlining check, and a generalized additive model with smooth terms fitted by restricted maximum likelihood in mgcv [53].

3.9 Software and reproducibility

The Python analysis used scikit-learn 1.8.0 [54], XGBoost 3.2.0, SHAP 0.51.0, statsmodels 0.15.0 and interpret 0.7.8. For the R replication, we used R 4.3.3 with gbm 2.1.8.1, ranger 0.16.0, mgcv 1.9.1, pROC 1.18.5, and haven 2.5.4. Random seeds were fixed in every script, at 20261001 in Python and at 7, 101 to 105 and 201 to 205 in R. A single script drew all figures from the saved results, so each plotted value can be traced to the analysis output.

4. Results and discussion

4.1 Descriptive statistics and attrition

The sample was predominantly female (93.1%), and 17.5% of respondents were only children (Table 2). Among the 875 students in the first-horizon sample, first-wave means were above the scale midpoints on every support, efficacy, and identity measure. Volitional identity had the lowest mean of the three identity facets (3.16), and the facets correlated between 0.75 and 0.78 with one another. Emotional support correlated 0.81 with learning support and 0.72 with ability support, so the support subscales overlap considerably. Variance inflation factors for the first-wave predictors in the logistic model ranged from 1.01 to 4.07, with the highest values for emotional support (4.07), learning support (3.47), and the identity facets (3.22 to 3.42). At the second wave, emotional support again had the largest value (5.76). This is the only value above 5, and none approaches 10, the level usually taken to signal serious collinearity. Even so, the overlap means that the logistic coefficients and the SHAP values of the support subscales should be read together. Students retained to the third wave reported slightly lower first-wave teacher support than those who dropped out (4.34 and 4.51, $p < 0.001$), and slightly lower teaching efficacy (4.12 and 4.20, $p = 0.033$). Their first-wave identity did not differ (3.29 and 3.33, $p = 0.155$). Attrition followed some predictors, but not the outcome itself, and Section 4.8 reports a check that weights for it.

4.2 Predictive performance

The three classifiers discriminated almost equally well (Table 5, Figure 3), and XGBoost's added value lies in the dependence plots, screening cut-points, and explanations for individual students that a single logistic slope per predictor does not provide. With first-wave features, the mean AUCs were 0.712 (SD 0.041) for XGBoost, 0.710 for the random forest, and 0.708 for logistic regression. Second-wave features improved all three, and XGBoost moved slightly ahead at 0.749 (SD 0.026), compared with 0.734 and 0.729. Under the corrected resampled t-test, its gain over logistic regression was not significant at either horizon (0.020 at the second, $t = 1.55$, $p = 0.134$, and 0.004 at the first, $p = 0.742$). XGBoost reached PR-AUCs of 0.468 and 0.503 at prevalences of 24.3% and 24.4%, indicating it placed later cases near the top of the ranking about twice as often as chance. In a label permutation test, the observed AUC of 0.706 compared with a mean of 0.499 and a maximum of 0.579 when labels were permuted ($p = 0.005$). Without the identity facets, the cold-start models reached AUCs between 0.668 and 0.686, 0.032 to 0.069 below the full models, which suggests that support and efficacy alone still separate later cases from the rest. Every model family did better at the second horizon, as expected for a shorter forecast.

Table 5. Cross-validated discrimination, accuracy, and calibration of the three classifiers in the four prediction tasks over 25 test folds

Task	Model	AUC (SD)	PR-AUC	Brier	Calibration slope	XGBoost minus LR in AUC (p)
Wave 1, full	Logistic regression	0.708 (0.041)	0.456	0.166	0.86	
	Random forest	0.710 (0.041)	0.473	0.164	0.83	
	XGBoost	0.712 (0.041)	0.468	0.164	0.86	+0.004 (0.742)
Wave 2, full	Logistic regression	0.729 (0.017)	0.477	0.162	0.88	
	Random forest	0.734 (0.030)	0.497	0.159	0.86	
	XGBoost	0.749 (0.026)	0.503	0.157	0.92	+0.020 (0.134)
Wave 1, cold start	Logistic regression	0.676 (0.041)	0.412	0.172	0.87	
	Random forest	0.668 (0.042)	0.398	0.173	0.77	
	XGBoost	0.677 (0.038)	0.411	0.172	0.80	+0.001 (0.940)
Wave 2, cold start	Logistic regression	0.686 (0.022)	0.410	0.171	0.89	
	Random forest	0.672 (0.023)	0.399	0.173	0.76	
	XGBoost	0.680 (0.024)	0.397	0.172	0.81	-0.006 (0.634)

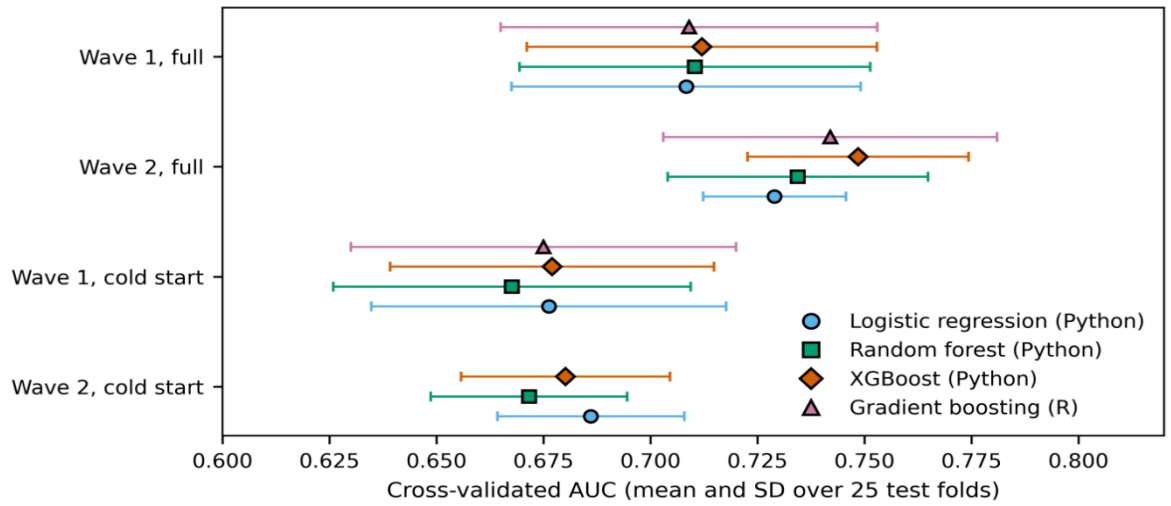


Figure 3. Cross-validated AUC for logistic regression, random forest, and XGBoost in Python, and gradient boosting in R, for the four prediction tasks. Markers show means, and bars show one standard deviation over 25 test folds. The R replication covered the three main tasks.

The first-wave models came close, however, and they leave more time to act. We next asked how much of this accuracy came from prior identity. A logistic regression on first-wave overall identity alone reached an AUC of 0.709, and one on second-wave overall identity reached 0.726. A simple yes-or-no indicator of prior weak identity reached 0.631 and 0.657. Compared with the identity baseline, the full XGBoost model gained 0.005 in AUC at the first horizon (p = 0.688) and 0.024 at the second (p = 0.082). Most of the ranking came from how strongly students already identified with teaching, while the other predictors widened the gap in predicted risk between later cases and other students, with a category-free NRI of 0.26 (95% CI 0.11 to 0.40) and an IDI of 0.032 (0.017 to 0.048) at the first horizon. Both were larger at the second horizon (NRI 0.38, 0.24 to 0.52, and IDI 0.050, 0.035 to 0.067). To see what support and efficacy contribute on their own, we added only these five subscales to prior identity. The AUC barely changed (0.003 with XGBoost, p = 0.815, and 0.002 with logistic regression, p = 0.853), but reclassification improved, with NRIs of 0.35 (0.19 to 0.50) and 0.19 (0.04 to 0.35). Support and efficacy refined each student's estimate with little change to who ranked highest, and they showed where that risk came from, which a single identity score cannot do.

Logistic regression kept pace with XGBoost throughout, consistent with the review by Christodoulou et al. [34].

4.3 Calibration

Observed rates tracked the predictions closely across risk deciles (Figure 4). XGBoost had a calibration slope of 0.86 at the first horizon and 0.92 at the second, with Brier scores of 0.164 and 0.157. Slopes are a little below 1.0, indicating mild overfitting, with the highest predicted risks slightly too extreme. Within this sample, the predicted probabilities can still be read as approximate risks without recalibration. Section 4.7's decision analysis relies on this property.

4.4 Global explanations

Volitional identity was the strongest predictor at the first horizon (Table 6, Figure 5). Its mean absolute SHAP value was 0.356 on the log-odds scale, about twice that of personal teaching efficacy (0.185). General teaching efficacy (0.174) and emotional support (0.169) followed closely. Gender, only-child status, and place of origin contributed little. Nearly half of all attribution went to the identity facets (44.7%), with smaller shares for efficacy (28.2%), support (22.6%), and the demographic indicators (4.6%).

Table 6. Mean absolute SHAP values on the log-odds scale, with Python and R ranks

Feature	Wave 1, full	Wave 2, full	Wave 1, cold start
Volitional identity	0.356 (1, R 1)	0.334 (1, R 1)	-
Personal teaching efficacy	0.185 (2, R 2)	0.150 (4, R 4)	0.289 (1, R 2)
General teaching efficacy	0.174 (3, R 4)	0.100 (6, R 8)	0.213 (3, R 3)
Emotional support	0.169 (4, R 2)	0.098 (8, R 6)	0.282 (2, R 1)
Extrinsic value identity	0.134 (5, R 5)	0.123 (5, R 5)	-
Intrinsic value identity	0.079 (6, R 6)	0.239 (2, R 2)	-
Learning support	0.064 (7, R 8)	0.099 (7, R 7)	0.111 (4, R 4)
Ability support	0.054 (8, R 7)	0.164 (3, R 3)	0.049 (5, R 5)
Only child	0.025 (9, R 10)	0.012 (9, R 9)	0.043 (6, R 6)
Female	0.018 (10, R 9)	0.002 (11, R 10)	0.019 (7, R 7)
Place of origin, code 2	0.015 (11, R 11)	0.005 (10, R 10)	0.011 (8, R 8)

Note: Ranks in brackets are from the Python XGBoost model and, after R, from the independent R gradient boosting model with sampled Shapley values. A dash marks a feature not used in that task.

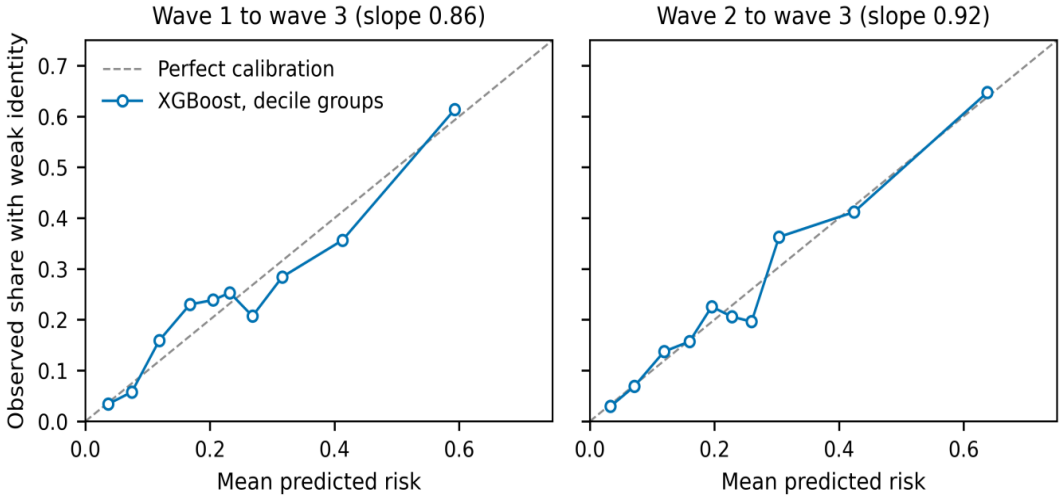


Figure 4. Calibration of the XGBoost out-of-fold predictions at the two forecast horizons. Each point compares the mean predicted risk with the observed share of weak identity in one decile of predicted risk, and the dashed line marks perfect calibration.

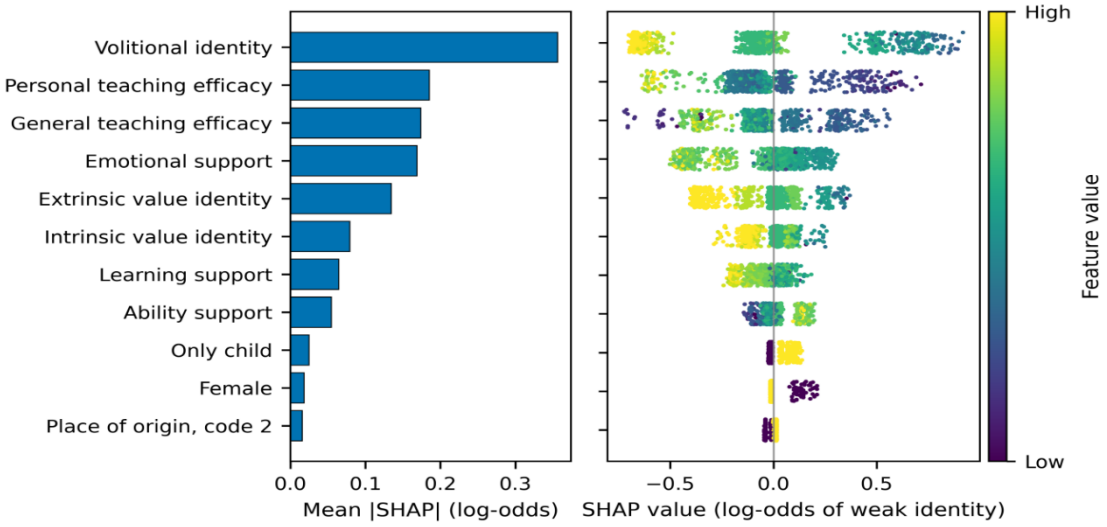


Figure 5. Global SHAP explanation of the wave 1 full model. Left, mean absolute SHAP value of each predictor on the log-odds scale. Right, SHAP value of each student for each predictor, colored from low to high feature value.

At the second horizon, the identity share rose to 52.6%, and intrinsic value, identity, and ability support moved up the ranking. In the cold-start model, personal teaching efficacy (0.289) and emotional support (0.282) shared the lead. Without identity items, the model relied on the two resources most closely tied to identity in earlier work [12,19]. Volitional identity ranked first in 15 of the 20 bootstrap refits, and the mean Kendall's tau between bootstrap and full-sample rankings was 0.71. The independent R analysis found the same leading feature in all three tasks, and its rankings agreed with the Python rankings (Kendall's tau of 0.88, 0.88, and 0.93). The lead of volitional identity fits the view that commitment, the willingness to keep choosing the profession, is central to identity and to decisions to stay or leave [5,19]. Personal teaching efficacy also outranked general teaching efficacy. This agrees with the view of Tschannen-Moran and Hoy [26] that general teaching efficacy is the less coherent of the two constructs, and with the close link between self-efficacy and occupational commitment reported by Klassen and Chiu [25].

4.5 Thresholds and non-linear effects

In the dependence plots of Figure 6, the main predictors change direction at identifiable points. Volitional identity had a step-shaped effect. Scores below 3.0 raised the predicted log-odds by about 0.56 to 0.70. Scores between 3.0 and 3.4 were close to neutral, while scores of 3.6 and above lowered the log-odds by about 0.6. Observed rates followed the same step, with weak identity reported later by 48.0% of the 204 students below 3.0 and by 17.1% of the 671 at or above it. Personal teaching efficacy changed from raising to lowering risk at about 3.64 on its six-point scale. Below that point, 47.8% of 138 students later reported weak identity, compared with 19.9% of students above it. These thresholds are the points at which smoothed running means of the SHAP values cross zero, and they are specific to this sample. Across the 20 bootstrap refits, the crossing for volitional identity stayed between 2.80 and 3.40 (median 3.00), and the crossing for personal teaching efficacy stayed between 3.55 and 3.91 (median 3.64).

Emotional support behaved less regularly, contributing most to risk between 3.5 and 4.0 and turning protective at about 4.5 in the full-sample model, after which the curve flattened. Students below 4.5 had a 30.8% rate of later weak identity, and those above it 12.4%. Unlike the two thresholds above, this one moved a great deal between bootstrap refits, from 3.17 to 4.50 (median 3.17). The learning support crossing, near 4.9 in the full sample, ranged from 3.44 to 4.89. For the support subscales, the curves show that risk fell as support rose, but they do not fix a single cut-off. Cubic splines improved the in-sample fit of the logistic model (likelihood-ratio $\chi^2 = 53.5$, df = 32, p = 0.010) but did not improve its out-of-sample AUC (0.699).

In R, a generalized additive model kept most smooth terms close to linear on the log-odds scale, including volitional identity and emotional support (effective degrees of freedom 1.00 for both), and showed curvature mainly for personal teaching efficacy (edf 3.53, p = 0.066). The step in the boosted curve for volitional identity therefore reflects the piecewise form of shallow trees, since its effect stayed close to linear on the log-odds scale. Explainable boosting machines reached AUCs of 0.720 without pairwise interactions and 0.716 with them, close to XGBoost's 0.714 and logistic regression's 0.708 in this comparison. These comparisons indicate that a linear model captured what mattered for prediction, and the steps in the SHAP curves serve mainly as practical operating points for screening. General teaching efficacy also had an irregular curve, but only 38 students sat at its low end. Ability support received small positive contributions at high values. Given its correlation of 0.72 with emotional support, this likely reflects how the model splits shared information between correlated features. Interaction terms accounted for 27.6% of the total SHAP interaction mass in the boosted model. General teaching efficacy with emotional support formed the strongest pair, followed by intrinsic value identity with volitional identity.

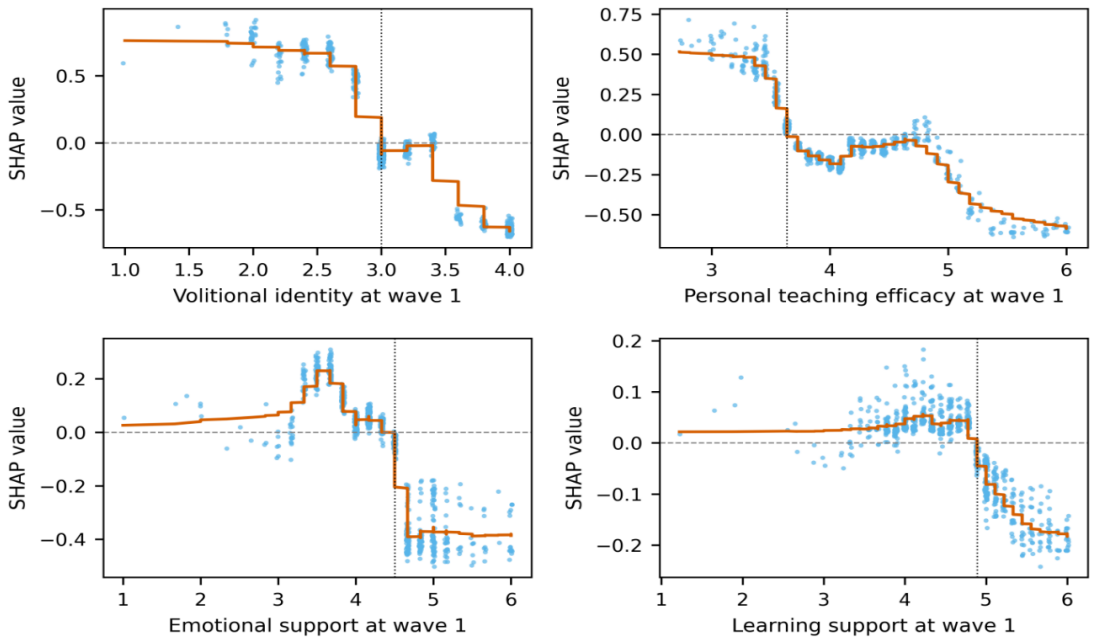


Figure 6. SHAP dependence plot for the Wave 1 full model of volitional identity, personal teaching efficacy, emotional support, and learning support. Points are students; lines show running means of the SHAP values, and dotted vertical lines mark where the running mean first crosses zero in the full-sample model.

Among the 183 students below both the volitional identity and the emotional support thresholds, 50.3% later reported weak identity. This share describes how the boosted model distributed its attributions; however, it did not translate into better discrimination. Adding the two strongest products to the logistic model changed neither its fit (χ^2 = 0.05, df = 2, p = 0.973) nor its AUC (0.708), and the same test in R gave p = 0.973. With XGBoost and logistic regression performing so similarly, the structure appears nearly additive on the log-odds scale. The direction of the strongest pair is still of interest, although the logistic test did not confirm it. In the boosted model, high emotional support lowered predicted risk further when general teaching efficacy was also high (mean interaction contribution -0.082 on the log-odds scale, 199 students). For students with high emotional support but low general teaching efficacy, the interaction raised predicted risk (0.107, 108 students). Self-determination theory links emotional support to relatedness, the sense that teachers care about the student [21]. Bandura's account adds that encouragement leads to effort only when the student expects that effort to pay off [22], and in teaching, that expectation is what general teaching efficacy measures [23,24]. Feeling cared for may therefore strengthen identity most when a student also believes that teaching can overcome the difficulties children bring from home. This reading remains a hypothesis that designed studies would need to test.

4.6 Individual explanations

To show what an individual explanation looks like, Figure 7 breaks down the forecast for the student with the highest out-of-fold predicted risk at the first horizon who later reported weak identity. Starting from a baseline log-odds of -1.15, her volitional identity score of 2.40 added 0.74, and her personal teaching efficacy of 3.27 added 0.51. Smaller contributions came from extrinsic value identity (2.33, adding 0.35), general teaching efficacy (adding 0.33), and emotional support (3.67, adding 0.24). Together with the remaining predictors, these produced a log-odds of 1.28, a predicted probability of 0.78. A tutor would begin with her commitment and her confidence as a teacher and turn to classroom support afterward, an order that a bare risk score could not suggest [38].

4.7 Decision utility and screening capacity

At each horizon, the models gave positive net benefit for referral thresholds from 0.10 to 0.50, above the options of referring everyone or no one (Figure 8). At a threshold of 0.30, the first-horizon model had a net benefit of 0.052, while referring everyone gave -0.081. Among every 100 students, the model referred 12.0 who later had weak identity and 15.9 who did not. A net benefit of 0.052 corresponds to about five correctly referred students per 100, once each unnecessary referral is weighted by the odds of the 0.30 threshold. At the second horizon, XGBoost had a higher net benefit than logistic regression at most thresholds above 0.25 (0.072 against 0.056 at 0.30). If the 20% of students with the highest first-wave risk were seen, 48.6% of them would be true cases against a base rate of 24.3%, and they would include 39.9% of all later cases (Table 7, Figure 9).

With second-wave data, the same budget would reach 43.4% of later cases at 52.9% precision. A 10% budget pushed precision above 60% at both horizons. Because precision depends on the base rate, these figures apply to cohorts with a similar share of weak identity. In such cohorts, a tutorial conversation costs little and does little harm when a flag is wrong, so precision at this level is enough to start one.

4.8 Robustness and replication

Straightlining flagged at the first wave affected 368 of 875 students (42.1%), and Table 8 lists this and the other checks. Without the flagged students, the AUC was 0.720 for XGBoost and 0.714 for logistic regression, and at the second horizon, 0.766 and 0.748. At the four alternative cut-offs from 2.8 to 3.2, prevalence ranged from 11.8% to 62.5%, the AUC of XGBoost stayed between 0.732 and 0.767, and the identity facets led every time. With the three-level ordinal outcome, out-of-fold concordance was 0.731. For the continuous score, cross-validated R^2 was 0.257 for linear regression and 0.227 for XGBoost. Without dichotomizing, the signal survives, although the cut-off still decides which students a program would call weak.

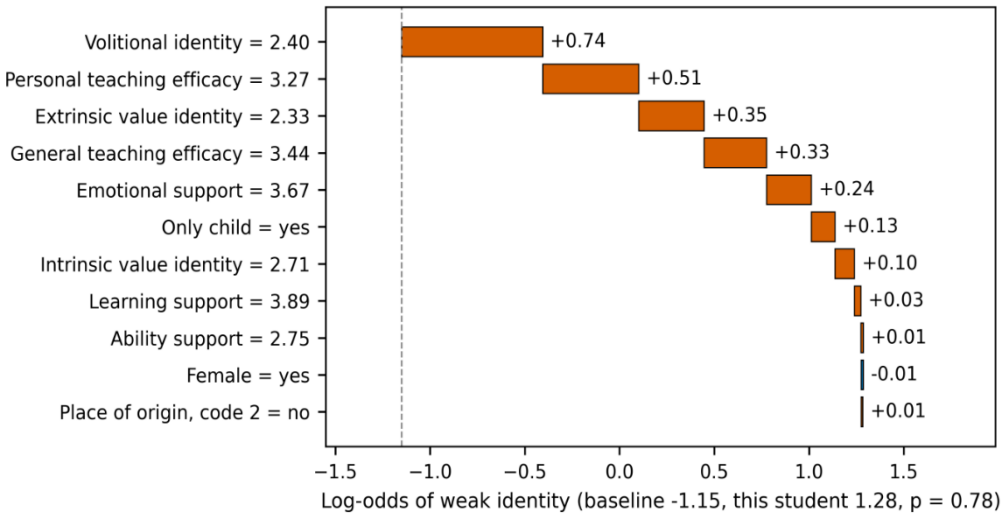


Figure 7. SHAP explanation of the first-wave forecast for the student with the highest out-of-fold predicted risk who later had weak identity. Each bar shows how much one predictor moved her predicted log-odds from the sample baseline. Orange bars raise the predicted risk and the blue bar lowers it.

Table 7. Screening capacity of the XGBoost model at fixed counseling budgets

Share flagged	Wave 1: flagged	Wave 1: cases captured (%)	Wave 1: precision (%)	Wave 2: flagged	Wave 2: cases captured (%)	Wave 2: precision (%)
10%	88	25.4	61.4	102	26.5	64.7
20%	175	39.9	48.6	204	43.4	52.9
30%	263	51.6	41.8	306	58.2	47.4
40%	350	60.1	36.6	408	66.3	40.4

Note: Students were ranked by out-of-fold predicted risk averaged over five repetitions. The base rates were 24.3% and 24.4%.

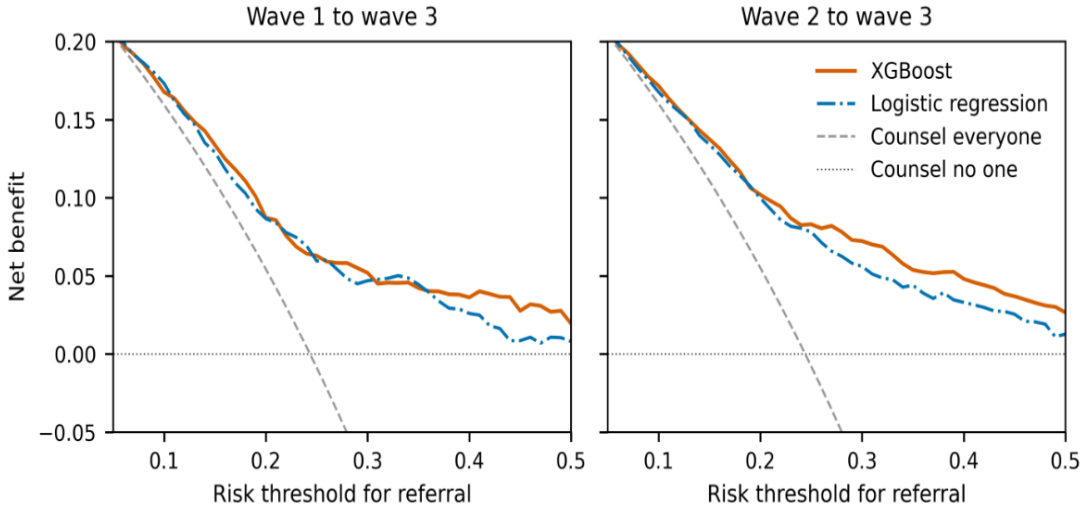


Figure 8. Decision curves for referral to counseling at the two forecast horizons, comparing XGBoost and logistic regression with referring every student and referring none. Net benefit counts true referrals minus unnecessary referrals, weighted by the odds at each risk threshold.

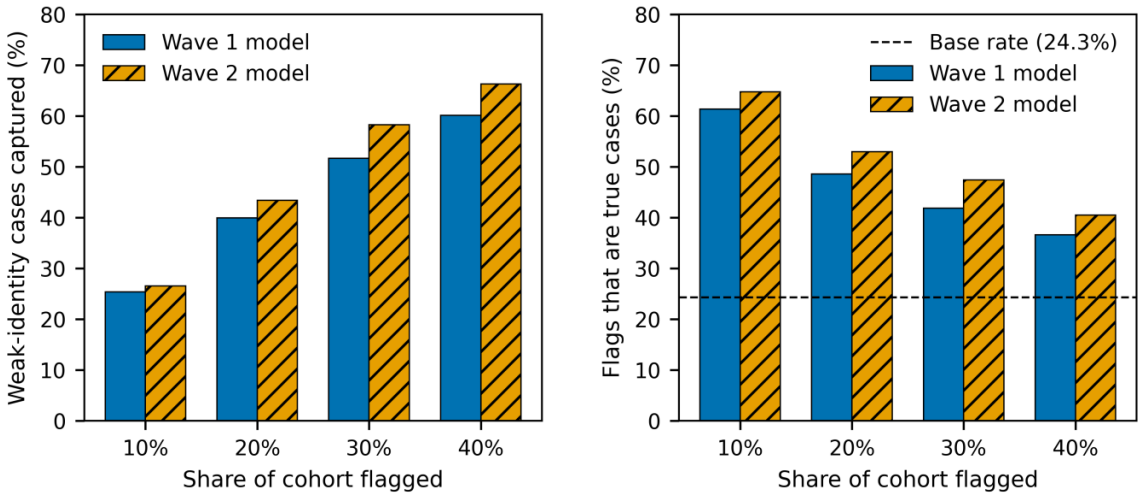


Figure 9. Screening capacity of the XGBoost model when a fixed share of the cohort is flagged. Left, share of later weak-identity cases captured. Right, share of flagged students who later had weak identity, with the dashed line marking the base rate.

Table 8. Baselines, robustness checks, and alternative specifications, mainly for the first-horizon full model

Check	**n**	**Prevalence (%)**	**XGBoost**	Comparison **model**	**p**
Main analysis, wave 1 full	875	24.3	0.712	0.708	0.742
Baseline: first-wave overall identity only (LR)	875	24.3	0.714 (full)	0.709	0.688
Baseline: first-wave weak status only (LR)	875	24.3	-	0.631	-
Baseline: second-wave overall identity only (LR)	1,020	24.4	0.750 (full)	0.726	0.082
Full model against baseline, NRI and IDI (wave 1)	875	24.3	NRI 0.26	IDI 0.032	-
Wave 1 overall identity plus support and efficacy (XGBoost)	875	24.3	0.712	0.709 (baseline)	0.815
Overall identity plus support and efficacy against baseline, NRI and IDI (wave 1)	875	24.3	NRI 0.35	IDI 0.027	-
Straightlining at wave 1, flagged respondents removed	507	29.6	0.720	0.714	0.658
Straightlining at wave 2, flagged respondents removed (wave 2 model)	487	32.0	0.766	0.748	0.117
Outcome includes a mean of exactly 3.0	875	57.5	0.747	0.741	0.447
Cut-off moved to 2.8	875	11.8	0.767	0.767	-
Cut-off moved to 2.9	875	18.1	0.732	0.734	-
Cut-off moved to 3.1	875	59.9	0.749	0.744	-
Cut-off moved to 3.2	875	62.5	0.755	0.755	-
Ordinal outcome, three levels (concordance)	875	-	-	0.731	-
Logistic regression with cubic splines	875	24.3	0.714	0.699	0.354
Logistic regression with two interactions	875	24.3	0.714	0.708	0.588
Explainable boosting machine, main effects	875	24.3	0.714	0.720	0.405
Explainable boosting machine, with pairs	875	24.3	0.714	0.716	0.707
Retention weights applied	875	24.3	0.713	0.708	0.711
Wave 1 and wave 2 features combined	674	24.6	0.741	0.733	-
Label permutation test (200 permutations)	875	24.3	0.706	null mean 0.499	0.005
Leave-cohort-out, wave 1 model, held-out cohort code 1	385	28.6	0.704	0.715	-
Leave-cohort-out, wave 1 model, held-out cohort code 2	488	21.1	0.723	0.709	-
Leave-cohort-out, wave 1 model, two cohorts pooled	873	24.4	0.685	0.677	-
Leave-cohort-out, wave 2 model, held-out cohort code 1	306	28.8	0.743	0.724	-
Leave-cohort-out, wave 2 model, held-out cohort code 2	371	21.3	0.771	0.746	-
Leave-cohort-out, wave 2 model, held-out cohort other codes	343	23.9	0.752	0.740	-
Leave-cohort-out, wave 2 model, three cohorts pooled	1,020	24.4	0.736	0.716	-
Continuous outcome, cross-validated R^2	875	-	0.227	0.257	-

Table 9. Subgroup audit of the XGBoost out-of-fold predictions

Model	**Subgroup**	**n**	**Cases**	**AUC**	**Mean predicted risk (%)**	**Observed rate (%)**	**Flagged at 20% budget (%)**
Wave 1	Male	58	17	0.720	25.7	29.3	27.6
	Female	817	196	0.713	24.1	24.0	19.5
	Not an only child	738	175	0.700	24.0	23.7	19.1
	Only child	137	38	0.783	25.6	27.7	24.8
	Place of origin, code 1	227	52	0.766	23.0	22.9	18.5
	Place of origin, code 2	648	161	0.696	24.7	24.8	20.5
Wave 2	Male	62	15	0.857	24.5	24.2	25.8
	Female	958	234	0.741	24.3	24.4	19.6
	Not an only child	853	206	0.750	24.4	24.2	20.6
	Only child	167	43	0.743	23.8	25.7	16.8
	Place of origin, code 1	302	67	0.745	21.7	22.2	13.9
	Place of origin, code 2	718	182	0.750	25.4	25.3	22.6

Retention weighting left the AUC at 0.713, and combining first-wave and second-wave features raised it to 0.741. XGBoost did not differ significantly from logistic regression in any of these checks. In leave-cohort-out validation, the first-horizon XGBoost model reached 0.704 and 0.723 in the two held-out cohorts, and the second-horizon model 0.743 to 0.771, close to the random-fold estimates. Pooled across cohorts, the values were lower (0.685 and 0.736), because the cohorts differed in their rates of weak identity, and a model trained on some cohorts shifts the predicted risks of another. In R, gradient boosting gave 0.713 and 0.726 in the held-out first-horizon cohorts. For the three main tasks, the R AUCs (0.709, 0.742, and 0.675) were within 0.007 of the Python values (0.712, 0.749, and 0.677).

4.9 Subgroup audit

The subgroup audit found no group for which the model was clearly miscalibrated (Table 9). AUC ranged from 0.696 to 0.783 across subgroups for the first-horizon model and from 0.741 to 0.857 for the second. Mean predicted risk stayed within 3.6 percentage points of the observed rate in every subgroup. Under a 20% budget, the share flagged varied with the observed rate. Male students, for example, were flagged more often at the first horizon (27.6% against 19.5%), matching their higher observed rate of weak identity (29.3% against 24.0%). With only 58 and 62 male students in the two samples, however, these estimates are imprecise. Any deployment should monitor calibration separately for male students. Future validation studies should also oversample male students or model gender as a moderator, in addition to reporting results by gender. The same caution applies to place of origin, whose two codes appear in the deposit without labels.

4.10 Implications for practice and research

Teacher educators could use the model in two steps, beginning at the start of a program with a short survey of support, efficacy, and identity to identify the students at highest risk. Where identity is not measured, five subscales, three for teacher support and two for teaching efficacy, are enough for the cold-start model to reach an AUC of about 0.68, which makes it the cheapest place to begin. For each flagged student, the SHAP explanation then indicates whether a tutor should first talk about commitment, confidence in teaching, or the student's relationship with staff. The thresholds found here are simple enough to use without the model. They are volitional identity below 3.0, personal teaching efficacy below about 3.6, and, less firmly, emotional support below about 4.5. Because they come from smoothed curves in one sample, and the support threshold moved widely between refits, each program should re-estimate them on its own data. For the student in Figure 7, three of the contributions translate directly into questions a tutor could ask.

- Her low volitional identity (2.40) added most to her risk. A tutor might ask what would make her want to keep choosing kindergarten teaching after graduation, and what makes her doubt it at present.
- Confidence in teaching came next, from a personal teaching efficacy of 3.27. She might be asked to recall a lesson that went well and to plan one small, observable teaching task for her next placement.
- Emotional support (3.67) added less, but it may be the easiest to act on. The tutor could ask which staff she feels able to approach and arrange a named contact for the coming term.

Calibration often worsens when a model moves to a new setting [36], and a program adopting this one should recalibrate it on its first cohort before relying on its risk estimates. Errors of educational prediction models can also fall unevenly on groups of students [41]. A flag is best used to open a supportive conversation and stay out of the student's academic record. Tutors are more likely to act on an explanation given in plain terms, such as the two or three features that raised a student's risk most [38]. For research on professional identity, the results go beyond the linear findings reviewed in Section 2. Prior identity carried most of the information about later identity, and studies of identity change need to model this persistence explicitly. Personal teaching efficacy predicted later identity more closely than general teaching efficacy. The boosted model gave about a quarter of its attributions to interactions, yet interaction terms did not improve a logistic model, so the joint effect of support and efficacy remains a hypothesis for designed studies. Machine learning studies in education can complete the checks asked for by current reporting guidance [40] on open data, from a persistence baseline to a subgroup audit and a second-package replication.

The cut-off of 3.0 can also be read against Hong's [5] account of professional identity, which describes value, efficacy, commitment and emotions as its components. On the four-point scale used here, a mean below 3.0 means that a student did not, on average, agree with statements about valuing teaching, wanting to stay in it, and acting on that choice. On this reading, the cut-off sits where commitment is no longer endorsed. Beijaard et al. [3] and Canrinus et al. [19] both treat commitment as a core indicator of identity, giving the cut-off substantive meaning in addition to its statistical role. Many students chose exactly 3.0, perhaps using the agree option as a default, but the cut-off sweep showed that the main conclusions held whether these students were counted as weak or not.

5. Conclusion

Using open three-wave data from 1,646 pre-service early childhood teachers in China, this study showed that weak professional identity can be forecast one or two waves ahead with useful and well-calibrated accuracy, and that an independent analysis in R reproduces the result. Logistic regression did almost as well, and prior identity alone came close, so the value of the boosted model lies mainly in its explanations and in the sharper individual risk estimates shown by the reclassification measures. Commitment, personal teaching efficacy and emotional support carried the forecasts. Where identity is not measured, a short survey of support and efficacy can stand in. With the full first-wave model, a program that screens one student in five would reach roughly two in five of those later at risk. Earlier work on pre-service teachers' identity reported average associations, and this study adds forecasts for individual students tested against counseling budgets a program can afford. Several limitations qualify these conclusions. Because the deposited file holds subscale means only, measurement reliability remains unchecked. Because the deposit omits the time between waves, the practical lead time of each forecast is unknown. The cut-off still has to be validated against a later career outcome, which requires data on intention to leave. All respondents come from early childhood teacher education in China and are predominantly female, so the thresholds may not transfer to other programs without re-estimation. The model was validated internally, across random folds and across cohorts, but external validation in other programs

remains to be done. Retention to the third wave was related to some predictors, although weighting did not change the results. SHAP values describe the model's predictions, and causal effects of support or efficacy on identity would need a different design. Any counseling triggered by the model would still require its own evaluation of effectiveness. Before the screen is used to allocate scarce counseling resources, the recommended next step is a prospective, multi-site validation with thresholds and decision rules registered in advance. Future work could also add placement and course data as predictors and test whether model-triggered counseling improves identity at graduation.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically with regard to authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with policies on research ethics. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The data analyzed in this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.20790669 under a CC BY 4.0 license. Additional data will be provided by the corresponding author upon reasonable request.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] Y. Yang, N. Rao, and J. Sun, "From pre-service preparation to professional development: Early childhood teachers' learning experiences, ECE quality, and child development in China," Early Education and Development, vol. 35, no. 5, pp. 1080-1102, 2024, doi: 10.1080/10409289.2024.2336659.

[2] Q. Ma, "A critical realist explanation of Chinese early childhood teachers' professional identity," Early Childhood Education Journal, vol. 54, no. 7, pp. 6281-6291, 2026, doi: 10.1007/s10643-026-02306-5.

[3] D. Beijaard, P. C. Meijer, and N. Verloop, "Reconsidering research on teachers' professional identity," Teaching and Teacher Education, vol. 20, no. 2, pp. 107-128, 2004, doi: 10.1016/j.tate.2003.07.001.

[4] C. Beauchamp and L. Thomas, "Understanding teacher identity: an overview of issues in the literature and implications for teacher education," Cambridge Journal of Education, vol. 39, no. 2, pp. 175-189, 2009, doi: 10.1080/03057640902902252.

[5] J. Y. Hong, "Pre-service and beginning teachers' professional identity and its relation to dropping out of the profession," Teaching and Teacher Education, vol. 26, no. 8, pp. 1530-1543, 2010, doi: 10.1016/j.tate.2010.06.003.

[6] D. J. Madigan and L. E. Kim, "Towards an understanding of teacher attrition: A meta-analysis of burnout, job satisfaction, and teachers' intentions to quit," Teaching and Teacher Education, vol. 105, Art. no. 103425, 2021, doi: 10.1016/j.tate.2021.103425.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4765-4774. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[8] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56-67, 2020, doi: 10.1038/s42256-019-0138-9.

[9] A. K. V. Olsen, I. G. H. Kjær, and K. S. Spieler, "Predictors of professional identity among early childhood teacher students," Education Sciences, vol. 14, no. 6, Art. no. 589, 2024, doi: 10.3390/educsci14060589.

[10] Y. Dai, X. Yang, and D. Li, "The formation of pre-service ECE teachers' professional identity: The roles of self-efficacy, learning motivation, and engagement," Acta Psychologica, vol. 265, Art. no. 106625, 2026, doi: 10.1016/j.actpsy.2026.106625.

[11] Y. Liu and Z. Xin, "Emotional labor and professional identity among pre-service early childhood teachers: the mediating role of emotional exhaustion and the moderating role of social and emotional competence," Frontiers in Psychology, vol. 17, Art. no. 1830373, 2026, doi: 10.3389/fpsyg.2026.1830373.

[12] X. Tian, M. Liu, and T. Yue, "Understanding professional identity through policy and support perceptions: A latent profile study of pre-service preschool teachers in China," Education Sciences, vol. 16, no. 7, Art. no. 1069, 2026, doi: 10.3390/educsci16071069.

[13] K. Ann and N. Boonnam, "A framework for predicting preservice teacher performance using machine learning and SMOTE-based class imbalance handling," Discover Education, vol. 5, no. 1, Art. no. 765, 2026, doi: 10.1007/s44217-026-02058-9.

[14] T. Yue, "Dataset for 'Professional identity development among preservice early childhood teachers: A cross-lagged panel network analysis of teacher self-efficacy and perceived teacher support'," Zenodo, 2026, doi: 10.5281/zenodo.20790669.

[15] F. Rodrigues and M. J. Mogarro, "Student teachers' professional identity: A review of research contributions," Educational Research Review, vol. 28, Art. no. 100286, 2019, doi: 10.1016/j.edurev.2019.100286.

[16] M. Izadinia, "A review of research on student teachers' professional identity," British Educational Research Journal, vol. 39, no. 4, pp. 694-713, 2013, doi: 10.1080/01411926.2012.679614.

[17] C. Lamote and N. Engels, "The development of student teachers' professional identity," European Journal of Teacher Education, vol. 33, no. 1, pp. 3-18, 2010, doi: 10.1080/02619760903457735.

[18] F. Hanna, R. Oostdam, S. E. Severiens, and B. J. H. Zijlstra, "Domains of teacher identity: A review of quantitative measurement instruments," Educational Research Review, vol. 27, pp. 15-27, 2019, doi: 10.1016/j.edurev.2019.01.003.

[19] E. T. Canrinus, M. Helms-Lorenz, D. Beijaard, J. Buitink, and A. Hofman, "Self-efficacy, job satisfaction,

motivation and commitment: exploring the relationships between indicators of teachers' professional identity," European Journal of Psychology of Education, vol. 27, no. 1, pp. 115-132, 2012, doi: 10.1007/s10212-011-0069-2.

[20] X. Ren, R. Feng, Z. Yan, J. Liu, and Z. Guo, "The relationship between occupational stigma consciousness and occupational identity among pre-service early childhood education teachers in China: a four-year longitudinal study," Frontiers in Psychology, vol. 17, Art. no. 1793509, 2026, doi: 10.3389/fpsyg.2026.1793509.

[21] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being," American Psychologist, vol. 55, no. 1, pp. 68-78, 2000, doi: 10.1037/0003-066X.55.1.68.

[22] A. Bandura, "Self-efficacy: Toward a unifying theory of behavioral change," Psychological Review, vol. 84, no. 2, pp. 191-215, 1977, doi: 10.1037/0033-295X.84.2.191.

[23] S. Gibson and M. H. Dembo, "Teacher efficacy: A construct validation," Journal of Educational Psychology, vol. 76, no. 4, pp. 569-582, 1984, doi: 10.1037/0022-0663.76.4.569.

[24] G. Yu, T. Xin, and J. Shen, "Teacher's sense of teaching efficacy: Its structure and influencing factors," Acta Psychologica Sinica, vol. 27, no. 2, pp. 159-166, 1995 (in Chinese). [Online]. Available: https://journal.psych.ac.cn/xlxb/CN/Y1995/V27/I2/159

[25] R. M. Klassen and M. M. Chiu, "The occupational commitment and intention to quit of practicing and pre-service teachers: Influence of self-efficacy, job stress, and teaching context," Contemporary Educational Psychology, vol. 36, no. 2, pp. 114-129, 2011, doi: 10.1016/j.cedpsych.2011.01.002.

[26] M. Tschannen-Moran and A. W. Hoy, "Teacher efficacy: capturing an elusive construct," Teaching and Teacher Education, vol. 17, no. 7, pp. 783-805, 2001, doi: 10.1016/S0742-051X(01)00036-1.

[27] A. J. Bowers, R. Sprott, and S. A. Taff, "Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity," The High School Journal, vol. 96, no. 2, pp. 77-100, 2013, doi: 10.1353/hsj.2013.0000.

[28] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," WIREs Data Mining and Knowledge Discovery, vol. 10, no. 3, Art. no. e1355, 2020, doi: 10.1002/widm.1355.

[29] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[30] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001, doi: 10.1214/aos/1013203451.

[31] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," Information Fusion, vol. 81, pp. 84-90, 2022, doi: 10.1016/j.inffus.2021.11.011.

[32] S. Tao, H. Zhang, and L. Ding, "Explainable machine learning for sustainable education: Predicting college students' reliance on generative artificial intelligence," Acta Psychologica, vol. 265, Art. no. 106622, 2026, doi: 10.1016/j.actpsy.2026.106622.

[33] G. Słowiński, Y. Kapranov, and B. Iwanowska, "Threshold-like structure and barrier-benefit asymmetry in students' behavioural intention toward e-learning: An explainable machine learning analysis in post-pandemic higher education," Sustainability, vol. 18, no. 10, Art. no. 4633, 2026, doi: 10.3390/su18104633.

[34] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," Journal of Clinical Epidemiology, vol. 110, pp. 12-22, 2019, doi: 10.1016/j.jclinepi.2019.02.004.

[35] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," BMC Bioinformatics, vol. 7, Art. no. 91, 2006, doi: 10.1186/1471-2105-7-91.

[36] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," BMC Medicine, vol. 17, Art. no. 230, 2019, doi: 10.1186/s12916-019-1466-7.

[37] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," Medical Decision Making, vol. 26, no. 6, pp. 565-574, 2006, doi: 10.1177/0272989X06295361.

[38] H. Khosravi, S. Buckingham Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gašević, "Explainable artificial intelligence in education," Computers and Education: Artificial Intelligence, vol. 3, Art. no. 100074, 2022, doi: 10.1016/j.caeai.2022.100074.

[39] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019, doi: 10.1038/s42256-019-0048-x.

[40] G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, Art. no. e078378, 2024, doi: 10.1136/bmj-2023-078378.

[41] R. S. Baker and A. Hawn, "Algorithmic bias in education," International Journal of Artificial Intelligence in Education, vol. 32, no. 4, pp. 1052-1092, 2022, doi: 10.1007/s40593-021-00285-9.

[42] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," Knowledge and Information Systems, vol. 41, no. 3, pp. 647-665, 2014, doi: 10.1007/s10115-013-0679-x.

[43] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.
[44] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, Art. no. e0118432, 2015, doi: 10.1371/journal.pone.0118432.
[45] C. Nadeau and Y. Bengio, "Inference for the generalization error," Machine Learning, vol. 52, no. 3, pp. 239-281, 2003, doi: 10.1023/A:1024068626366.
[46] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values," Artificial Intelligence, vol. 298, Art. no. 103502, 2021, doi: 10.1016/j.artint.2021.103502.
[47] M. J. Pencina, R. B. D'Agostino Sr, R. B. D'Agostino Jr, and R. S. Vasan, "Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond," Statistics in Medicine, vol. 27, no. 2, pp. 157-172, 2008, doi: 10.1002/sim.2929.
[48] M. J. Pencina, R. B. D'Agostino Sr, and E. W. Steyerberg, "Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers," Statistics in Medicine, vol. 30, no. 1, pp. 11-21, 2011, doi: 10.1002/sim.4085.
[49] Y. Kim, J. Dykema, J. Stevenson, P. Black, and D. P. Moberg, "Straightlining: Overview of measurement, comparison of indicators, and effects in mail-web mixed-mode surveys," Social Science Computer Review, vol. 37, no. 2, pp. 214-233, 2019, doi: 10.1177/0894439317752406.
[50] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions," in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 623-631, doi: 10.1145/2487575.2487579.
[51] S. R. Seaman and I. R. White, "Review of inverse probability weighting for dealing with missing data," Statistical Methods in Medical Research, vol. 22, no. 3, pp. 278-295, 2013, doi: 10.1177/0962280210395740.
[52] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," Ecography, vol. 40, no. 8, pp. 913-929, 2017, doi: 10.1111/ecog.02881.
[53] S. N. Wood, "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," Journal of the Royal Statistical Society Series B: Statistical Methodology, vol. 73, no. 1, pp. 3-36, 2011, doi: 10.1111/j.1467-9868.2010.00749.x.
[54] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html



This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).