JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

An Explainable Counterfactual Decision-Support Framework for Early Academic Risk Prediction in Learning Analytics Systems

Jarot Budiasto [1,*], Gerzon Jokomen Maulany [1], Tatik Melinda Tallulembang [1], and Susanto [2]

[1] Department of Information Systems, Universitas Musamus, Merauke 99611, Indonesia; e-mail : jarot@unmus.ac.id; maulany@unmus.ac.id; tatik_melinda@unmus.ac.id
[2] Department of Informatics Engineering, Universitas Musamus, Merauke 99611, Indonesia; e-mail : susanto@unmus.ac.id
* Corresponding Author : Jarot Budiasto

Abstract: This study addresses the prediction-to-action gap in learning analytics by introducing an explainable counterfactual decision-support framework for translating early academic risk predictions into actionable and plausibility-aware intervention recommendations. The framework integrates gradient-boosted prediction, Shapley-value interpretation, actionability constraints, calibrated nearest-neighbor plausibility assessment, robust proximity evaluation, and personalized intervention mapping within a human-in-the-loop workflow. Experiments were conducted using the Open University Learning Analytics Dataset at day-30, day-60, and day-90 cutoffs, with group-based, leakage-controlled temporal splitting. The validation-optimized model achieved F1-scores of 0.6898, 0.7137, and 0.7306, respectively, while the Brier score decreased from 0.1842 to 0.1448. For 120 targeted enrolments (representing 119 distinct students) per cutoff, the generator produced 360 counterfactuals, with mean predicted-risk reduction increasing from 0.4543 to 0.6101 while maintaining stable robust proximity, at an average sparsity of approximately 4.7 changed features. Favorable-neighbor anchoring, rather than random search, enabled plausible recourse to be generated broadly. Robustness analyses showed that probability calibration did not change validity, while independently trained evaluators agreed with the direction of risk reduction for up to 98.3% of counterfactuals, although they estimated smaller effects from the same changes. These findings indicate that the direction of academic recourse is relatively stable across model families, whereas the magnitude required to cross an alert threshold remains estimator-dependent. Accordingly, the recommendations provide model-informed behavioral guidance rather than fixed intervention targets and require real-world validation. A controlled comparison with reproduced DiCE and NICE baselines under a common protocol retained recourse for every targeted student at calibrated plausibility, a result not achieved by three of the four baselines, while delivering approximately twice their predicted-risk reduction at a higher computational cost.

Keywords: Academic Risk Prediction; Algorithmic Recourse; Counterfactual Explanation; Decision Support Systems; Early Warning Systems; Educational Data Mining; Explainable Artificial Intelligence; Learning Analytics.

Received: July, 28th 2026
Revised: November, 18th 2026
Accepted: November, 19th 2026
Published: November, 22nd 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) licenses (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

Academic information systems, learning management systems, and virtual learning environments (VLEs) have made student learning behaviors increasingly observable through assessment submissions, online interactions, resource-access patterns, and academic outcomes. For higher education institutions, these data extend beyond routine records of learning activities and represent valuable resources for decision support and educational business intelligence. Accordingly, learning analytics and educational data mining have become important approaches for identifying students who may require academic assistance before final grades or withdrawal decisions occur [1]–[4]. Machine learning techniques have been widely

applied to predict student performance, dropout, and academic risk by integrating demographic information, registration records, engagement indicators, and assessment-related behaviors [5]–[8].

However, the effectiveness of early academic risk prediction should not be assessed solely as a classification problem. From an information systems perspective, the practical value of predictive analytics depends on whether its outputs can be translated into useful information for academic advisers, study programs, and institutional decision-makers. This creates an inherent trade-off in early-warning systems. Predictions made later in a course can benefit from more extensive behavioral and assessment evidence, potentially improving predictive performance, but may leave insufficient time for preventive intervention. Earlier predictions provide greater opportunities for academic support but must operate under more limited evidence and greater uncertainty. Consequently, early-warning systems need to be evaluated not only in terms of predictive performance but also according to their ability to provide interpretable, reliable, and actionable information for academic decision-making [4], [7]–[11].

Explainable artificial intelligence (XAI) has consequently received increasing attention in educational applications, where predictive outputs need to be understandable to teachers, academic advisers, students, and institutional decision-makers [12]–[15]. SHapley Additive exPlanations (SHAP) provide a means of identifying the features contributing most strongly to individual risk predictions, making complex model outputs more transparent [16], [17]. Recent OULAD-based explainable learning analytics studies have also demonstrated the usefulness of SHAP in identifying engagement and assessment features associated with student success and academic risk [18]–[20]. Nevertheless, feature-importance explanations primarily describe the factors associated with a prediction; they do not directly indicate which realistic changes could potentially reduce a student's predicted risk. This limitation has motivated the exploration of explanation approaches incorporating causal structure, including the combination of local interpretable models and structural causal reasoning in educational settings [21], [22].

Counterfactual explanations offer a complementary approach by describing how a model prediction would change under alternative input-feature values [23]–[28]. Recent studies have begun applying counterfactual explanations to student success prediction, dropout support, and educational decision-making [29]–[32]. In academic risk prediction, for example, a counterfactual may indicate that a student classified as being at risk could receive a lower predicted risk if learning activity increased, assignments were submitted earlier, or early assessment performance improved. However, the usefulness of such recommendations depends on more than their ability to change a model prediction. Counterfactuals must also be actionable and plausible, while respecting ethical constraints. In particular, immutable characteristics such as gender, disability, age, and prior education should not be presented as intervention targets. Without appropriate constraints, a mathematically valid counterfactual may therefore produce recommendations that are unrealistic, unattainable, or inappropriate for academic intervention.

Despite advances in academic risk prediction and explainable learning analytics, translating predictive explanations into actionable intervention recommendations remains insufficiently addressed. Existing studies have explored counterfactual recourse for student success and dropout support, but several methodological challenges remain when these approaches are considered in an early-warning decision-support context. First, counterfactual generation is often evaluated at a single observation point, providing limited insight into how recourse can support interventions across different stages of a course. Second, the use of default proximity objectives does not necessarily ensure that generated counterfactuals remain consistent with the observed data distribution, highlighting the need for empirically calibrated, data-driven plausibility assessment. Third, evaluating counterfactuals only with the model that generated them leaves uncertain whether the recommended changes remain effective when assessed by independently trained model families. These challenges motivate a framework that jointly considers temporal early-warning design, SHAP-informed feature prioritization, risk-stratified student selection, calibrated nearest-neighbor plausibility, robust proximity, and intervention mapping.

To address these challenges, this study proposes a plausibility-aware counterfactual decision-support framework with SHAP-informed feature prioritization for early academic risk prediction in learning analytics systems. The framework integrates temporally staged risk

prediction and counterfactual generation with ethically constrained actionability, data-driven plausibility calibration, and cross-model robustness assessment within a human-in-the-loop decision-support workflow. Rather than claiming that the individual techniques are entirely new, this study focuses on their systematic integration and evaluation under a multi-cutoff OULAD early-warning setting. Specifically, the research aims to develop a temporal, leakage-controlled risk prediction model across day-30, day-60, and day-90 cutoffs; generate ethically constrained and plausibility-aware counterfactual explanations; and translate counterfactual feature changes into personalized academic intervention categories. The main contributions are as follows:

- A risk-stratified counterfactual decision-support framework that supports early academic risk assessment across multiple temporal cutoffs.
- Ethically constrained actionability rules that distinguish feature mutability and incorporate directional constraints into counterfactual generation.
- An empirically calibrated nearest-neighbor plausibility assessment combined with robust proximity scaling to prevent inflated distances for near-zero-variance activity features.
- An ablation study isolating the effect of favorable-neighbor anchoring on counterfactual coverage and diversity.
- A cross-model and calibration robustness analysis that distinguishes the consistency of recourse direction from its estimator-dependent magnitude.
- A mapping between SHAP-based risk drivers and operational intervention categories to support academic advising workflows.

The remainder of this paper is organized as follows. Section 2 reviews related work on academic risk prediction, explainable learning analytics, and counterfactual recourse, and positions this study against the most closely related approaches. Section 3 describes the proposed framework, including temporal feature engineering, model development, actionability constraints, and the plausibility-aware counterfactual generator. Section 4 presents the experimental results and discussion, covering predictive performance, counterfactual quality, favorable-neighbor anchoring ablation, cross-model robustness, and study limitations. Finally, Section 5 concludes the paper and outlines directions for future work.

2. Related Work

Across this literature, the Open University Learning Analytics Dataset (OULAD) has become a standard public benchmark because it links demographics, registration, virtual learning environment activity, assessments, and final results in a single relational structure [33], enabling time-aware prediction while demanding careful prevention of temporal leakage. Tree-based gradient boosting suits such heterogeneous tabular data and integrates naturally with tree-based explanation methods [34]. However, prior evaluations have largely stopped at discrimination metrics. Probabilistic calibration, subgroup robustness, and the ethical considerations surrounding human oversight and student vulnerability receive comparatively little attention [35]–[37], even though these properties determine whether a risk score can responsibly drive an institutional intervention.

Counterfactual explanations complement feature-importance methods by describing how a prediction would change under alternative feature values [23]–[28]. Desirable counterfactuals are simultaneously valid, proximate, sparse, plausible, diverse, and actionable, and recent educational studies have begun to explore them for student success and dropout support [29]–[32]. These studies also expose the central difficulty: without actionability and plausibility constraints, a technically valid counterfactual may recommend an unattainable or ethically inappropriate change, such as altering a student's prior education. From an information systems perspective, institutions increasingly seek to convert integrated academic data into planning and data-driven decisions [9]–[11], [38], which requires explanations that can be operationalized as interventions rather than merely inspected.

The closest studies address complementary aspects of educational counterfactual explanations. Afrin et al. [32] examined individual DiCE explanations for social-media-based student-success prediction and complemented the analysis with an educator user study. Buñay-Guisñan et al. [30] extended counterfactual reasoning to clustered student groups using OULAD clickstreams, whereas Hidayat et al. [39] combined longitudinal cohort validation,

temporal-velocity features, SHAP attribution, and constrained DiCE on a large institutional dataset. Table 1 summarizes their methodological positioning relative to the present framework.

Table 1. Positioning against representative education-focused counterfactual studies

Dimension	Afrin et al. [32]	Buñay-Guisñan et al. [30]	Hidayat et al. [39]	This study
Temporal design	Static pass/fail prediction with a random 80/20 split	Full-course weekly click-streams aggregated into static totals	Longitudinal prediction with LOCO historical-to-future-cohort validation	Day-30/60/90 early-warning cutoffs with group-based splitting
Feature representation and guidance	Social-media usage features; no separate attribution stage	Aggregated click features with K-means grouping and DiCE-derived key-feature selection	Temporal-velocity feature engineering with SHAP attribution	Temporal features with SHAP-guided prioritization of actionable variables
Actionability constraints	Age, gender, and WAM fixed; usage variables allowed to vary	Click features only and treated as actionable; no explicit immutable-feature taxonomy	Forward-looking behavioral features with a three-tier causal constraint architecture	Immutable, semi-actionable, and actionable feature groups with directional constraints
Plausibility mechanism	DiCE resemblance and diversity with basic immutability constraints	Warren group-CF algorithm with proximity and sparsity evaluation	Causal constraint architecture; no calibrated NN threshold reported	Empirical range checks and LOO-calibrated nearest-neighbor threshold with robust scaling
Recourse unit and targeting	Individual failing students; three alternatives	Cluster-level group recourse with individual DiCE as comparator	Individual borderline high-risk students; up to four alternatives	Individual recourse stratified into three risk levels; up to three alternatives
Decision-support translation	Individual what-if explanations evaluated by educators	Group-level remedial guidance without implemented intervention evaluation	Behavior-oriented intervention pathways	Intervention-category mapping within a human-in-the-loop workflow
Counterfactual validation scope	Single-model generation with an 18-educator user study	Group-versus-individual comparison of coverage, number of changes, and runtime	Single-model constrained-counterfactual feasibility and risk-reduction analysis	Cross-model directional agreement, calibration-sensitivity checks, and anchoring ablation
Data setting	Social-media usage and academic-outcome data	OULAD DDD-course click-stream data	Large longitudinal institutional records with LMS and academic-progression features	OULAD multi-module relational learning-analytics data

Note: Dimensions summarize procedures explicitly reported in each study rather than overall quality or relative superiority; the absence of a calibrated nearest-neighbor threshold does not imply the absence of feasibility or plausibility controls.

As shown in Table 1, prior studies provide important advances in individual explanation, group recourse, educator-oriented validation, and longitudinal prescriptive analytics. The present study does not replace these contributions, but combines multi-cutoff early-warning prediction, SHAP-guided actionable feature selection, empirically calibrated nearest-neighbor plausibility, risk-stratified recourse, and cross-model evaluation within a unified human-in-the-loop decision-support workflow.

3. Proposed Method

3.1. Research Design

This study adopted an experimental information-systems and learning-analytics design to develop a plausibility-aware counterfactual decision-support framework for early academic risk prediction. The framework is positioned as an analytical component that can support academic information systems or learning analytics dashboards, rather than as an autonomous decision maker. It connects three analytical stages: risk prediction, SHAP-based interpretation, and counterfactual-based intervention planning. The workflow is presented in Figure 1.

The experiment began with OULAD relational tables, followed by temporal preprocessing and leakage prevention. Cutoff-specific datasets were constructed for days 30, 60, and 90, and data were split using a group-based train-validation-test design. Baseline models were compared before eXtreme Gradient Boosting (XGBoost) was selected and threshold-tuned on validation data. SHAP was used to identify risk drivers, while actionability constraints and risk stratification guided plausibility-aware counterfactual generation. Counterfactual quality evaluation then assessed validity, risk reduction, robust proximity, and calibrated nearest-neighbor plausibility before recommendations were mapped to personalized academic intervention suggestions.

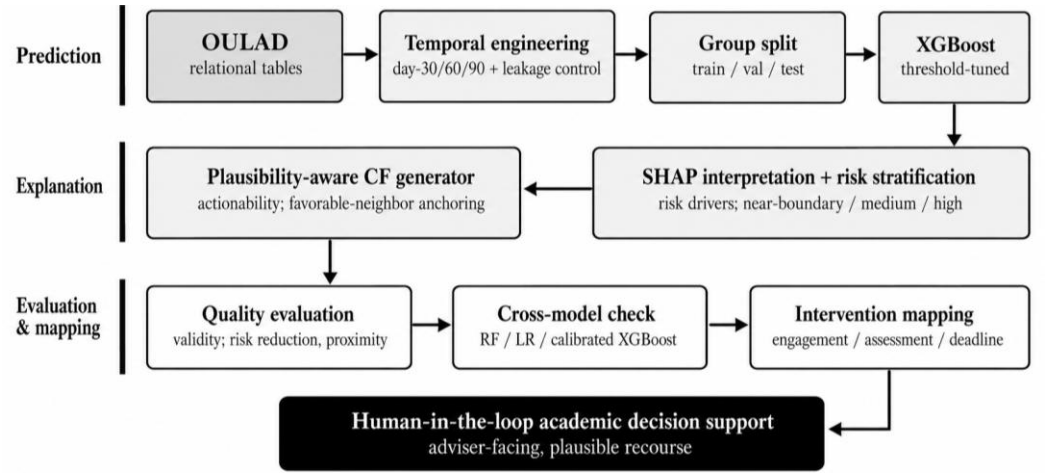

Figure 1. Workflow of the plausibility-aware counterfactual decision-support framework, connecting temporal prediction, SHAP-based explanation, and counterfactual-driven intervention mapping

3.2. Dataset and Target Definition

The study used OULAD, which contains demographic, registration, VLE, assessment, and outcome records for students in online higher education courses [33]. The original outcome variable was transformed into a binary academic risk label: Fail and Withdrawn were labelled as at risk, while Pass and Distinction were labelled as not at risk. Three temporal prediction scenarios were evaluated: day 30, day 60, and day 90, representing early, middle, and later stages of the course. The at-risk label groups Fail and Withdrawn because both represent unsuccessful course outcomes that early-warning support aims to prevent, whereas Pass and Distinction represent successful completion; this binary framing reflects the operational goal of flagging students who will not complete the course successfully. Only records available on or before each cutoff day were used to construct that cutoff's features, so that no post-cutoff information could leak into prediction. Numerical features with missing values were imputed using the training-set median, and all preprocessing statistics were estimated on the training split only. To prevent student-level leakage, records were grouped by student identifier so that a given student appeared in only one of the training, validation, or test partitions. Table 2 summarizes the cutoff-specific datasets.

Table 2. Cutoff-specific dataset summary

Cutoff	Observations	Attributes	At-risk rate	Not at risk
30	27,466	52	0.4399	15,385
60	26,361	53	0.4164	15,385
90	25,562	53	0.3981	15,385

Note: Attributes refer to variables before categorical encoding.

3.3. Temporal Feature Engineering and Leakage Prevention

Temporal feature engineering was performed separately for each cutoff day. VLE activity features were aggregated only from events observed on or before the cutoff. Assessment features were computed only from assessment records whose relevant dates and submissions were available no later than the cutoff. Students who had already withdrawn before or on the

cutoff day were excluded to prevent temporal leakage. The generated features included demographic variables, registration indicators, VLE engagement features, assessment performance features, submission timing variables, and behavioral participation indicators. The date_unregistration field was excluded from modelling because it directly relates to withdrawal and may leak target information.

3.4. Model Development, Threshold Selection, and SHAP Interpretation

Four classifiers were evaluated before selecting the main model: Dummy Classifier, Logistic Regression, Random Forest, and XGBoost. XGBoost was included because gradient boosting has been widely used for structured tabular prediction and supports efficient tree-based explanation through SHAP [16], [17], [34] and has been optimized for tabular prediction tasks in related work [40]. SHAP is used specifically to rank the actionable features by their global importance; this ranking then determines which actionable features are prioritized as candidates for the counterfactual search. SHAP therefore informs feature prioritization rather than driving the counterfactual optimization objective itself, which is governed by the plausibility, actionability, and proximity criteria described in Section 3.6. These methodological choices are aligned with the characteristics of OULAD and of the early academic-risk task. The data are tabular and heterogeneous, combining demographic, registration, engagement, and assessment features with non-linear interactions and frequent missing values, for which gradient-boosted trees are well suited and require limited feature scaling. SHAP provides consistent local and global attribution for tree ensembles, allowing the most influential actionable drivers to be prioritized for recourse. Because the behavioral feature vectors are dense and vary in scale, a nearest-neighbor plausibility criterion grounded in the observed training distribution offers a data-driven notion of feasibility, and favorable-neighbor anchoring ties each recommendation to real retained-student profiles rather than to synthetic points. Finally, because counterfactual recommendations are known to be sensitive to the choice of predictor, cross-model evaluation tests whether the recommended recourse remains directionally stable when the same changes are scored by independently trained models. Data splitting used id_student as the grouping key to prevent student-level leakage; the training set was used for model fitting, the validation set for threshold selection, and the independent test set for final evaluation. Thresholds from 0.10 to 0.90 were tested on the validation set, and the threshold that maximized F1-score was selected for each cutoff. For reproducibility, all models used a fixed random seed (42). The XGBoost generator was configured with 450 estimators, a maximum tree depth of 4, a learning rate of 0.05, and subsample and column-sample ratios of 0.85, using the binary logistic objective with the log-loss evaluation metric; class imbalance was handled through scale_pos_weight set to the training-set negative-to-positive ratio. The baseline models followed standard configurations: Logistic Regression with balanced class weights and a 1000-iteration limit, Random Forest with 300 trees, a minimum of two samples per leaf, and balanced-subsample weighting, and a most-frequent Dummy classifier as a naive reference. Threshold tuning was performed only on the validation split to avoid information leakage into the test set. Performance was assessed using accuracy, precision, recall, F1-score, ROC-AUC, PR-AUC, Brier score, and related confusion-matrix metrics. SHAP values were computed using a sample of up to 1,000 observations, and global feature importance was calculated using mean absolute SHAP values. Using the validation-selected threshold, the predicted class was obtained using Eq. (1), where $f(x_i)$ denotes the predicted academic risk probability for student i and τ denotes the selected threshold.

$$\hat{y}_i = \begin{cases} 1, & f(x_i) \geq \tau \\ 0, & f(x_i) < \tau \end{cases} \quad (1)$$

3.5. Actionability Constraints and Counterfactual Generation

Features were classified into immutable, semi-actionable, and actionable categories, as summarized in Table 3. Immutable features such as gender, disability, age band, region, prior education, and course module were not allowed to change. Semi-actionable features such as studied credits and registration timing were excluded from the main counterfactual search. Only actionable VLE and assessment features were allowed to change, with directional constraints applied to avoid pedagogically inappropriate recommendations.

Table 3. Feature actionability categories and treatment

Feature type	Examples	Counterfactual treatment
Immutable	gender, age band, disability, region, prior education, course module	Not modified
Semi-actionable	studied credits, previous attempts, registration timing	Excluded in main experiment
Actionable	VLE active days, assessment scores, assessment count, submission timing, resource access	Allowed with directional constraints

Counterfactual explanations were generated for students predicted as at risk. Students were stratified into near-boundary, medium-risk, and high-risk groups based on predicted risk score. Near-boundary students were those between the selected threshold and threshold + 0.20; medium-risk students were above that range and up to 0.80; high-risk students were above 0.80. The plausibility-aware generator combined random directional candidate search with favorable-neighbor anchoring, where favorable anchors were observed training profiles that were not at risk and predicted below the threshold. Candidate counterfactuals were ranked by validity, target achievement, calibrated nearest-neighbor plausibility, range plausibility, risk reduction, sparsity, and robust proximity, following counterfactual explanation literature emphasizing that valid counterfactuals should also be feasible, sparse, plausible, and actionable [23]–[28].

3.6. Plausibility and Robust Proximity Evaluation

A generated counterfactual instance was considered valid when it moved the predicted risk below the selected threshold, as defined in Eq. (2). Risk reduction was computed using Eq. (3), where the two terms denote the initial and counterfactual risk probabilities, respectively. Throughout this paper, risk reduction refers to a reduction in the model's predicted risk probability rather than an empirically observed change in a student's academic outcome; because the study does not track students who act on the recommendations, all reported reductions should be read as predicted-risk reductions rather than demonstrated causal effects.

$$V_i^{cf} = \begin{cases} 1, & f\left(x_i^{cf}\right) < \tau \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$\Delta r_i = f(x_i) - f\left(x_i^{cf}\right) \tag{3}$$

Sparsity was reported as the number of changed features. To avoid inflated distance values for sparse activity features with zero or near-zero IQR, robust normalized proximity was computed using Eq. (4),

$$P_i = \frac{1}{|C_i|} \sum_{j \in C_i} \frac{\left|x_{ij}^{cf} - x_{ij}\right|}{s_j}, s_j = max\{IQR_j,\ 0.5\left(q_{95,j} - q_{05,j}\right),\ 0.25\left(q_{99,j} - q_{01,j}\right),\ 1.0\} \tag{4}$$

where C_i denotes the set of changed features and s_j is the robust feature scale estimated from the training distribution.

The lower bound of 1.0 is retained as a numerical safeguard to prevent near-zero scaling values for features with very limited variation. This floor keeps the normalization denominator bounded and avoids disproportionately large proximity contributions from low-variance features. Calibrated nearest-neighbor plausibility was evaluated using Eq. (5). A counterfactual was considered nearest-neighbor plausible if its distance to the nearest at-risk training profile did not exceed the 99th percentile of leave-one-out nearest-neighbor distances computed within the same at-risk training population.

$$d_{\mathrm{NN}}\left(x_i^{\mathrm{cf}}\right) = \min_{z \in D_{\mathrm{train}}^{\mathrm{risk}}} d\left(x_i^{\mathrm{cf}}, z\right) \tag{5}$$

Here, $D_{\text{train}}^{\text{risk}}$ denotes the set of at-risk profiles in the training data, and $\delta_{0.99}$ represents the 99th percentile of leave-one-out nearest-neighbor distances computed within the same at-risk training population. Thus, the plausibility criterion evaluates whether a generated counterfactual lies within the calibrated neighborhood of observed at-risk training profiles.

A counterfactual was considered nearest-neighbor plausible if its distance to the nearest at-risk training profile did not exceed the 99th percentile of leave-one-out nearest-neighbor distances computed within the same at-risk training population. The 99th percentile was chosen as a conservative manifold boundary: it accepts recourse targets that lie within the range of all but the most extreme one percent of observed at-risk training profiles, thereby excluding implausible outliers while remaining permissive enough that feasible recourse can still be found for the targeted at-risk cases. This balances plausibility against attainability; a substantially lower percentile would risk rejecting realistic recommendations, whereas a higher cutoff would admit near-outlier targets. The stability of the downstream results supports this choice: the validation-selected prediction thresholds coincided with the peak F1 at each cutoff and F1 varied by only about 0.01 within a wide neighborhood of these thresholds, indicating that the pipeline is not finely tuned to a single operating point. To confirm that the results are not overly dependent on this choice, a sensitivity analysis was conducted by re-evaluating the generated counterfactuals under the 90th, 95th, and 99th plausibility percentiles. The recourse quality was found to be invariant to the threshold: mean risk reduction, sparsity, and robust proximity were identical across all three percentiles at every cutoff (for example, risk reduction of 0.4543, 0.5463, and 0.6101 at day 30, 60, and 90). The percentile affects only the nearest-neighbor acceptance rate, which declined from 100% at the 99th percentile to roughly 82% to 91% at the 95th and 60% to 74% at the 90th, as expected for a stricter manifold boundary. Because the substantive recourse metrics do not change, the 99th percentile governs how conservatively plausibility is enforced rather than what recommendations are produced, confirming that the reported findings are robust to this parameter.

4. Results and Discussion

4.1. Predictive Performance Across Temporal Cutoffs

The predictive model showed a consistent temporal pattern across the three cutoff scenarios. Table 4 reports the comparative performance of all four evaluated models at a common default threshold. XGBoost achieved the strongest discrimination and calibration at every cutoff, with the highest ROC-AUC (0.7912, 0.8341, 0.8638) and PR-AUC (0.7607, 0.8169, 0.8415) and the lowest Brier score, while the Dummy baseline (which never predicts the minority class) confirms that the task is non-trivial and that the learned models provide substantial improvement over a naive predictor. Logistic Regression and Random Forest were competitive but consistently below XGBoost on the ranking and calibration metrics, supporting its selection as the generator model. As shown in Table 4 and Figure 2, the validation-optimized XGBoost model achieved F1-scores of 0.6898, 0.7137, and 0.7306 at day 30, day 60, and day 90, respectively. ROC-AUC and PR-AUC improved over time, while the Brier score decreased from 0.1842 to 0.1448, indicating that predictive discrimination and probabilistic reliability improved as additional learning evidence became available.

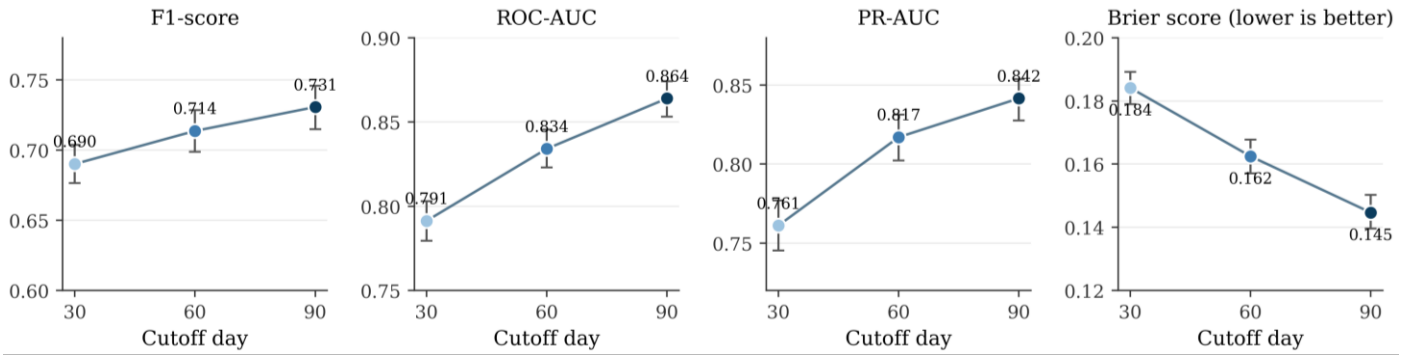

Figure 2. Validation-optimized XGBoost predictive performance across day-30, day-60, and day-90 cutoffs, with 95% bootstrap confidence intervals (1000 resamples)

Table 4. Comparative predictive performance of the evaluated models across cutoffs

Cutoff	Model	Acc.	Prec.	Recall	F1	ROC-AUC	PR-AUC	Brier
30	Dummy	0.5720	0.0000	0.0000	0.0000	0.5000	0.4280	0.4280
	Logistic Regression	0.6969	0.6364	0.6807	0.6578	0.7703	0.7278	0.1941
	Random Forest	0.7159	0.6968	0.5954	0.6421	0.7824	0.7480	0.1856
	XGBoost	0.7206	0.6739	0.6730	0.6735	0.7912	0.7607	0.1842
60	Dummy	0.5788	0.0000	0.0000	0.0000	0.5000	0.4212	0.4212
	Logistic Regression	0.7212	0.6501	0.7323	0.6887	0.8064	0.7685	0.1797
	Random Forest	0.7578	0.7540	0.6309	0.6870	0.8280	0.8105	0.1623
	XGBoost	0.7607	0.7218	0.7027	0.7121	0.8341	0.8169	0.1625
90	Dummy	0.6010	0.0000	0.0000	0.0000	0.5000	0.3990	0.3990
	Logistic Regression	0.7497	0.6745	0.7205	0.6967	0.8281	0.7911	0.1661
	Random Forest	0.7930	0.7894	0.6562	0.7167	0.8546	0.8318	0.1461
	XGBoost	0.7902	0.7438	0.7234	0.7335	0.8638	0.8415	0.1448

Note: Reported at the default 0.5 decision threshold for a fair cross-model comparison; validation-optimized XGBoost thresholds (0.42, 0.48, 0.53) are used for all subsequent counterfactual analyses.

4.2. SHAP-Based Risk Drivers

SHAP analysis identified assessment performance, VLE engagement, submission timing, and selected background or course-related variables as important risk drivers. Table 5 and Figure 3 show that assess_mean_score and vle_active_days were the two strongest global features. This is pedagogically meaningful because the strongest modifiable drivers relate to assessment remediation, engagement support, and deadline management.

Table 5. Top global SHAP features

Rank	Feature	Mean |SHAP|	Interpretation
1	assess_mean_score	0.3119	Average early assessment performance
2	vle_active_days	0.3116	Consistency of VLE engagement
3	code_module_GGG	0.2393	Course/module effect
4	assess_weighted_score_sum	0.2257	Cumulative weighted assessment evidence
5	assess_mean_days_before_due	0.1985	Submission timing
6	highest_education_Lower_Than_A_Level	0.1887	Prior educational background
7	assess_weighted_score_avg	0.1671	Weighted early assessment performance
8	assess_max_score	0.1416	Best early assessment evidence
9	vle_activity_page	0.1316	Use of page-based resources
10	vle_unique_sites	0.1170	Breadth of resource access

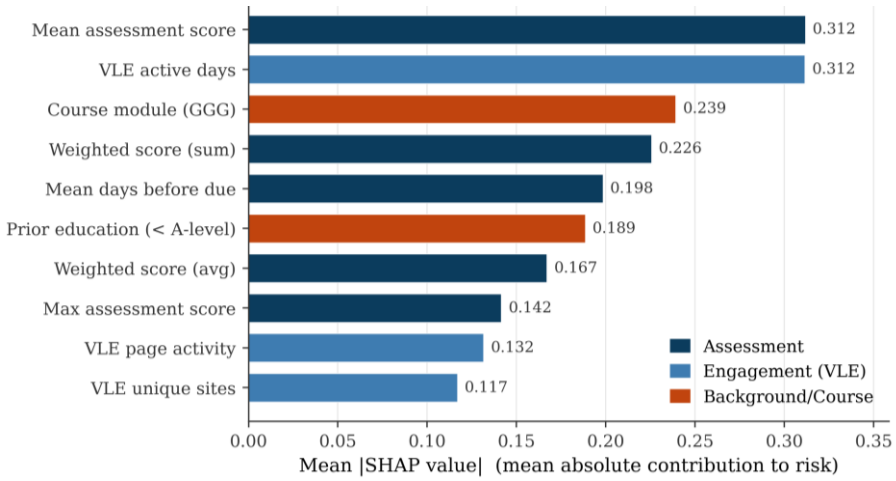

Figure 3. Global feature importance (mean absolute SHAP value) for early academic-risk prediction, grouped by feature family

4.3. Plausibility Calibration and Counterfactual Quality

Table 6 shows that the calibrated 99th-percentile plausibility thresholds ranged from 12.4338 to 13.9241 across cutoffs. The mean counterfactual nearest-neighbor distance remained below these thresholds in all scenarios, so every retained counterfactual satisfied the calibrated plausibility criterion.

Table 6. Leave-one-out nearest-neighbor plausibility calibration

Cutoff	Profiles	Actionable feat.	Q50 dist.	Q95 dist.	Q99 dist.
30	17,559	37	1.2725	6.9907	13.9241
60	16,851	38	1.6456	6.4408	12.4338
90	16,363	38	1.6898	6.4704	12.6598

Note: Distances are leave-one-out nearest-neighbor distances computed among at-risk training profiles; Q99 denotes the 99th-percentile plausibility threshold.

Table 7 summarizes the main counterfactual quality results. Within the risk-stratified selection design, each of the 120 targeted at-risk enrolments per cutoff received up to three counterfactuals, yielding 360 counterfactuals and 100% per-enrolment recourse coverage. The 120 enrolments were not sampled at random: they are the at-risk enrolments whose predicted risk fell within the recourse-eligible band at each cutoff, so that counterfactual generation was applied only to cases for which a meaningful risk reduction was both needed and potentially attainable. This targeted pool is therefore representative of borderline and higher-risk students rather than of the entire at-risk population, and the coverage results reported below should be read accordingly. These 120 enrolments correspond to 119 distinct students, because a small number of OULAD students are registered in more than one module presentation and are therefore evaluated separately in each. Because only candidates satisfying the validity, plausibility, and actionability constraints are retained by construction, these acceptance rates are necessarily equal to one and should be read as a feasibility guarantee of the generator rather than as an empirical discovery; the informative quantities are therefore coverage, risk reduction, sparsity, and robust proximity. It should be emphasized that the resulting coverage and validity of 100% are obtained on this curated target pool after plausibility and actionability filtering, and must not be interpreted as evidence that the framework generates valid counterfactuals for the entire population of at-risk students. More informatively, the retained counterfactuals do not merely satisfy the plausibility constraint but sit well inside the data manifold: the median counterfactual nearest-neighbor distance is only about 30% to 32% of the calibrated 99th-percentile threshold (4.51, 3.70, and 4.02 against thresholds of 13.92, 12.43, and 12.66 at day 30, 60, and 90), indicating that plausibility is achieved with a substantial safety margin rather than at the boundary of admissibility. As reported in Table 7, mean predicted-risk reduction increased steadily from the earliest to the latest cutoff while robust proximity remained stable, indicating that later cutoffs afford larger predicted-risk reductions without requiring more distant changes. These values should be interpreted relative to the imposed actionability constraints and the calibrated 99th-percentile nearest-neighbor threshold, not as evidence that the recommendations are causally guaranteed to improve academic outcomes.

Table 7. Plausibility-aware counterfactual quality metrics

Cutoff	Enrolments	CFs	Coverage	Mean risk red.	Mean sparsity	Robust prox.
30	120	360	100%	0.4543	4.6917	1.4135
60	120	360	100%	0.5463	4.7472	1.2026
90	120	360	100%	0.6101	4.7222	1.2229

Note: Enrolments denote targeted at-risk module registrations (120 per cutoff, corresponding to 119 distinct students); coverage is the proportion receiving at least one counterfactual. Four directional-constraint violations in the initial experiments are documented in the limitations.

Figure 4 visualizes the individual recourse trajectories. Each line traces one counterfactual from its original predicted risk to its counterfactual risk, coloured by risk stratum and faceted by cutoff. All trajectories terminate below the decision threshold (dashed line), and

the high-risk stratum exhibits the steepest descent, confirming at the level of individual students that the generated recourse moves every targeted case out of the at-risk region.

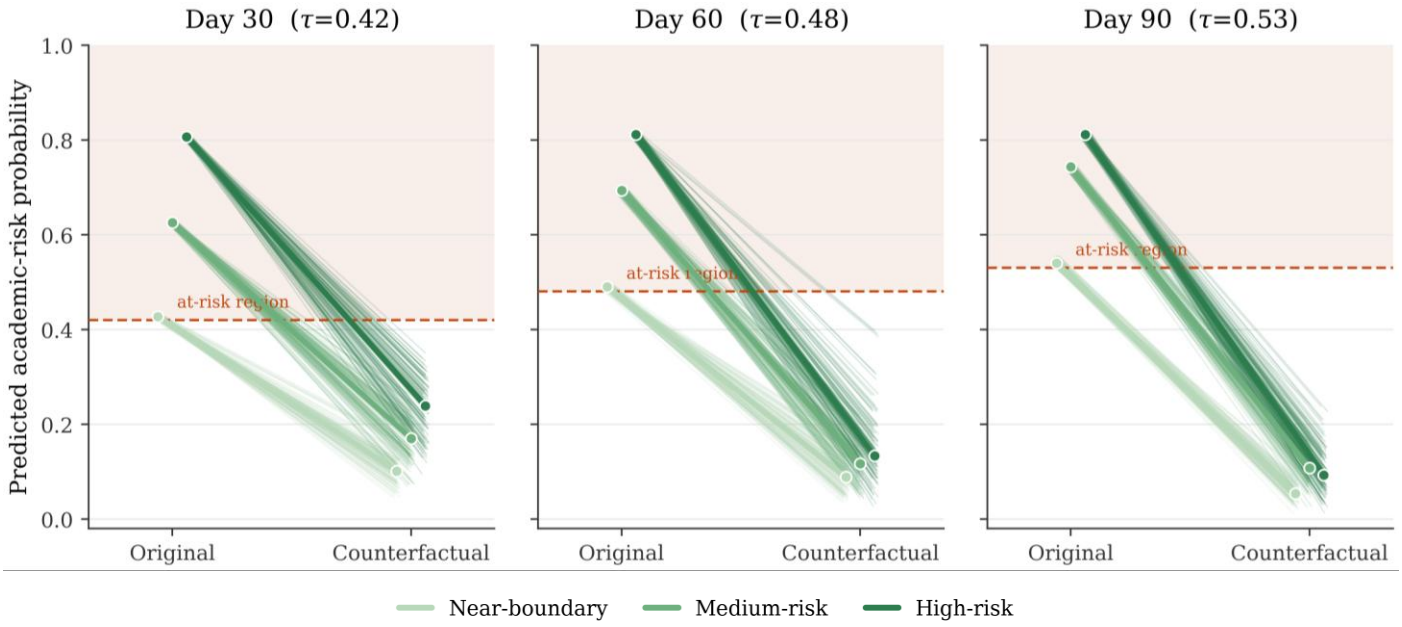

Figure 4. Individual counterfactual recourse trajectories from original to counterfactual predicted risk, by risk stratum and cutoff. The shaded band marks the at-risk region above the decision threshold; bold lines are per-stratum medians

Figure 5 summarizes these quality metrics across strata and cutoffs. Panel (a) shows that mean risk reduction rises monotonically from near-boundary to high-risk students and strengthens at later cutoffs, while panel (b) plots the per-counterfactual trade-off between robust proximity and risk reduction, with the cutoff centroids indicating stable proximity as risk reduction improves over time.

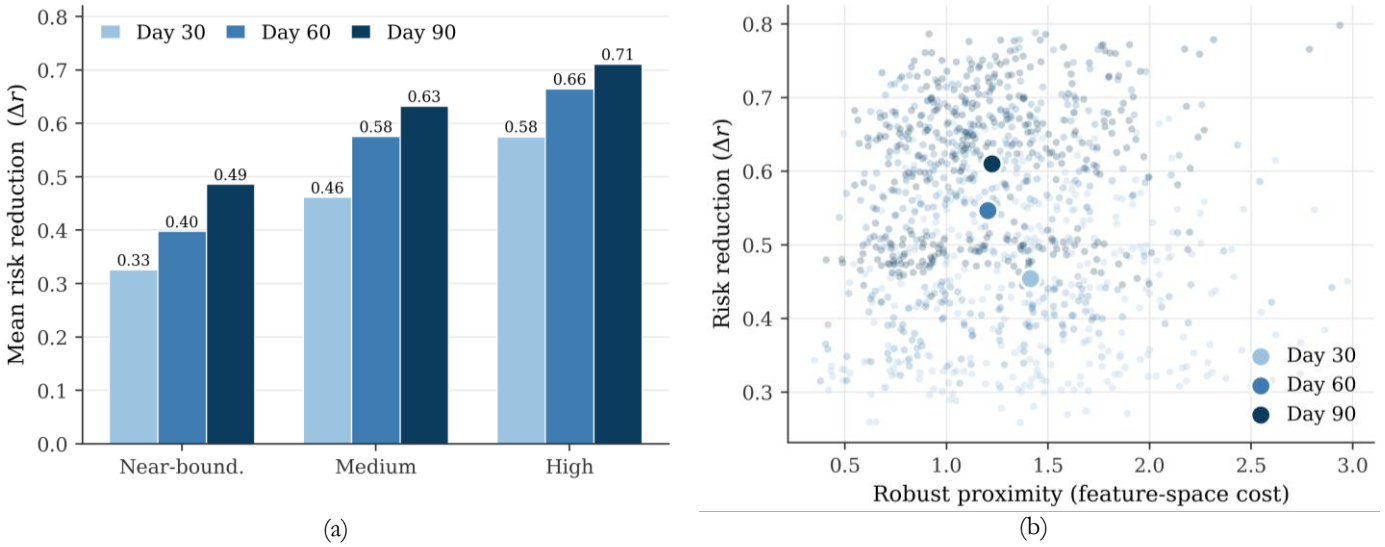

Figure 5. Counterfactual quality: (**a**) mean risk reduction by risk stratum and cutoff; (**b**) per-counterfactual cost-benefit of recourse (robust proximity versus risk reduction), with cutoff centroids

4.4. Counterfactual Results by Risk Stratum

Risk-stratified analysis showed that mean risk reduction increased from near-boundary to high-risk students, while sparsity remained relatively stable because the plausibility-aware search tended to generate multi-feature intervention packages. Table 8 shows that high-risk students obtained the largest average risk reduction, reaching 0.5752, 0.6648, and 0.7113 at day 30, day 60, and day 90, respectively, consistent with the larger gap these students must

close to fall below the threshold. Robust proximity did not vary monotonically with risk: at day 30 the medium-risk stratum required the most distant changes (1.4595), whereas at day 90 the high-risk stratum did (1.3451). This reflects the interaction between how far a student is from the decision boundary and how much reliable evidence is available at each cutoff; at earlier cutoffs with sparser behavioral evidence, the feasible favorable anchors for medium-risk students lie comparatively farther in the robust-scaled feature space. All strata retained full acceptance under the validity and plausibility constraints. The directional constraints were met in all but four counterfactuals at day 90, one in the near-boundary stratum and three in the high-risk stratum, which are identified in the seventh limitation and are included in the means reported here.

Table 8. Counterfactual quality by risk stratum

Cutoff	Risk stratum	Initial risk	CF risk	Risk red.	Sparsity	Robust prox.
30	near-boundary	0.4274	0.1014	0.3259	4.6333	1.4185
	medium-risk	0.6254	0.1637	0.4617	4.7000	1.4595
	high-risk	0.8069	0.2317	0.5752	4.7417	1.3623
60	near-boundary	0.4896	0.0913	0.3984	4.7667	1.2242
	medium-risk	0.6939	0.1183	0.5756	4.7250	1.2097
	high-risk	0.8116	0.1468	0.6648	4.7500	1.1738
90	near-boundary	0.5422	0.0558	0.4863	4.7500	1.1432
	medium-risk	0.7422	0.1095	0.6327	4.7667	1.1804
	high-risk	0.8112	0.0999	0.7113	4.6500	1.3451

Note: Values are averaged within each risk stratum; the constraints stated in the note to Table 7 apply here as well.

4.5. Ablation: Contribution of Favorable-Neighbor Anchoring

Table 9 reports the ablation separating retained counterfactuals by candidate source (random directional versus neighbor-anchored search). Neighbor anchoring supplied the large majority of retained counterfactuals (78.3%–83.3% across cutoffs) and, more importantly, reached almost every targeted student (119, 119, and 117 at day 30, 60, and 90), whereas random directional search produced admissible recourse for only 49, 53, and 57 students.

Table 9. Ablation of candidate source (neighbor anchoring vs random directional)

Cutoff	Source	CFs	Students reached	Diversity /student	Mean risk red.	Robust prox.
30	Neighbor anchor	300	119	2.47	0.4591	1.4145
	Random directional	60	49	1.22	0.4301	1.4082
60	Neighbor anchor	296	119	2.46	0.5506	1.2318
	Random directional	64	53	1.21	0.5261	1.0673
90	Neighbor anchor	282	117	2.38	0.6145	1.2155
	Random directional	78	57	1.35	0.5941	1.2499

Note: Students reached = distinct students with ≥1 retained counterfactual from that source; diversity/student = mean distinct counterfactuals per student.

Anchoring also roughly doubled the mean number of distinct counterfactuals per student (about 2.4 versus 1.2), improving recourse diversity, while achieving comparable or slightly higher risk reduction at similar robust proximity and sparsity. This indicates that the anchoring mechanism, rather than random perturbation, is what makes plausibility-constrained recourse broadly attainable; without it, roughly half of the at-risk students would receive no admissible recommendation. To further isolate the contribution of the individual search components beyond anchoring, a supplementary ablation compared three configurations of the generator under a common protocol: the full plausibility-aware search, a variant driven only by SHAP-prioritized features, and an unguided random actionable search. The full plausibility-aware search achieved perfect nearest-neighbor plausibility and validity (both 1.000) with the highest mean risk reduction (0.526). Removing the plausibility and anchoring machinery and relying on random actionable search degraded every dimension—validity fell

to 0.922, nearest-neighbor plausibility to 0.967, and mean risk reduction to 0.398, at worse robust proximity—while the SHAP-guided configuration recovered most of the quality (0.989 plausibility, 1.000 validity, 0.511 risk reduction), confirming that SHAP-based prioritization and the calibrated plausibility and anchoring components each contribute to the final recourse quality rather than any single one acting alone.

4.6. Actionable Features and Intervention Categories

The most frequently changed feature was vle_active_days, followed by vle_unique_sites, assess_mean_score, assess_weighted_score_sum, and assess_weighted_score_avg. Table 10 shows that the counterfactual generator mainly targeted engagement and assessment-related features, aligning with the SHAP results and supporting the interpretation that academic risk is shaped by engagement consistency, learning-resource breadth, and assessment performance. It is important to distinguish between two kinds of actionable features. Engagement features such as active days and the breadth of resources accessed are directly manipulable by students and correspond to behaviors they can change with appropriate support. Assessment-derived features, in contrast, are performance outcomes rather than actions a student can perform directly; a counterfactual that raises an assessment score should therefore be read as a target to be achieved through instructional support rather than as a self-directed action. Accordingly, changes in these features are mapped to staff-delivered support categories—for example, tutoring or assessment remediation—in which advisers or teachers help the student reach the indicated improvement, rather than implying that the student can simply increase a score at will. The resulting intervention categories were dominated by engagement support (2,642 changes) and assessment remediation or tutoring (1,879), followed by deadline management support (419), assessment support (114), and assessment participation support (44), indicating that plausibility-aware search with favorable-neighbor anchoring produced broader multi-feature intervention packages centered on engagement and assessment. These counts cover all 5,098 feature changes of the main counterfactual run and therefore include the four changes in a prohibited direction identified in the seventh limitation. The absence of such changes reported in Section 4.8 refers to the corrected constraint ordering used in the comparative and cross-model experiments, not to the run summarized here.

Table 10. Most frequently changed counterfactual features

Rank	Feature	Count	Intervention meaning
1	vle_active_days	813	Engagement consistency
2	vle_unique_sites	596	Broader resource access
3	assess_mean_score	591	Assessment remediation
4	assess_weighted_score_sum	497	Cumulative assessment evidence
5	assess_weighted_score_avg	410	Weighted assessment remediation
6	vle_activity_quiz	299	Quiz engagement
7	assess_mean_days_before_due	293	Deadline management
8	vle_activity_oucontent	239	Content engagement
9	assess_max_score	225	Assessment improvement
10	vle_event_count	216	Learning-event participation

4.7. Cross-Model and Calibration Robustness of Counterfactual Validity

Counterfactual validity is defined on the same XGBoost model used for generation, so a threshold crossing guarantees only a lower predicted probability. To test whether the recommended changes are model-specific artefacts, the reconstructed counterfactuals available for each cutoff were re-scored by two independently trained classifiers (Random Forest and Logistic Regression) and by the calibrated XGBoost probabilities (isotonic and Platt scaling fitted on the validation split). Each evaluator used its own validation-optimal threshold. Two complementary quantities are reported: cross-model validity, the share of counterfactuals falling below the evaluator's threshold, and risk-reduction agreement, the share for which the evaluator agrees that risk decreased at all. Table 11 and Figure 6 report the results for the 357 reconstructed counterfactuals per cutoff. These are the counterfactuals as generated: the four

that move a feature in a prohibited direction were not excluded before re-scoring, so the agreement rates below describe the generated set rather than an admissible subset of it.

Probability calibration left validity unchanged. Both isotonic and Platt scaling retained a validity rate of 1.0000 at every cutoff, because calibration is monotonic in the base-model score and the retained counterfactuals sit well below the decision threshold rather than grazing it, with a mean validity margin of 0.25 to 0.44 and, at day 90, every counterfactual at least 0.29 below the threshold. The concern that the reported validity is an artefact of imperfect calibration is therefore not supported.

Table 11. Cross-model and calibration robustness of counterfactual validity

Cutoff	Evaluator	Thr.	Cross-validity	Risk-red. agreement	Mean orig. risk	Mean CF risk
30	XGBoost (generator)	0.42	1.0000	1.0000	0.620	0.165
	XGBoost + isotonic	0.34	1.0000	1.0000	0.569	0.160
	XGBoost + Platt	0.37	1.0000	1.0000	0.591	0.155
	Logistic Regression	0.39	0.9804	0.9972	0.587	0.098
	Random Forest	0.38	0.3725	0.9384	0.572	0.403
60	XGBoost (generator)	0.48	1.0000	1.0000	0.664	0.118
	XGBoost + isotonic	0.34	1.0000	1.0000	0.592	0.096
	XGBoost + Platt	0.38	1.0000	1.0000	0.632	0.101
	Logistic Regression	0.49	0.7563	0.8291	0.646	0.259
	Random Forest	0.37	0.5658	0.9636	0.585	0.357
90	XGBoost (generator)	0.53	1.0000	1.0000	0.698	0.088
	XGBoost + isotonic	0.34	1.0000	1.0000	0.625	0.070
	XGBoost + Platt	0.45	1.0000	1.0000	0.659	0.068
	Logistic Regression	0.51	0.7059	0.7591	0.639	0.334
	Random Forest	0.45	0.7703	0.9832	0.590	0.370

Note: Each evaluator uses its own validation-optimal threshold. Cross-validity = share of counterfactuals below that threshold; risk-reduction agreement = share for which the evaluator agrees risk decreased.

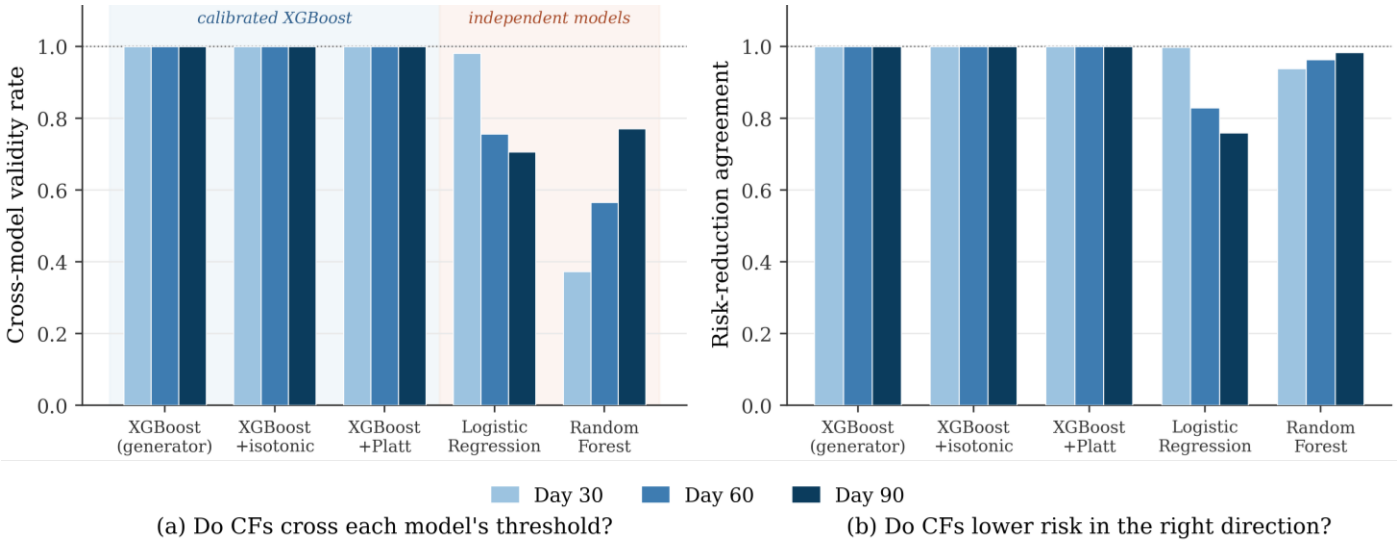

Figure 6. Cross-model robustness of counterfactual explanations: (**a**) validity rate based on each evaluator's decision threshold; (**b**) agreement on the direction of risk reduction across evaluators.

The independently trained evaluators show a more nuanced pattern that separates the direction of the recourse from its magnitude. Both models agree with the direction of change in the large majority of cases: Random Forest concurs that risk decreases for 93.8%, 96.4%, and 98.3% of counterfactuals at day 30, 60, and 90, and Logistic Regression for 99.7%, 82.9%,

and 75.9%. Cross-model validity, however, is lower for Random Forest (0.3725, 0.5658, 0.7703). The mechanism is visible in the last two columns of Table 11: on the identical feature changes, Random Forest reduces predicted risk by only 0.17 to 0.23 on average, whereas XGBoost reduces it by 0.46 to 0.61. Random Forest is a systematically more conservative estimator of the effect of the same intervention, so the recommended changes often move risk in the right direction without travelling far enough to cross its own threshold. Cross-model validity also improves monotonically with the cutoff day for Random Forest, consistent with later cutoffs providing more behavioural evidence and therefore more decisive feature changes.

The practical reading is that the direction of the recommended recourse is generally consistent across model families, although the interpretation should account for the four directional-constraint violations retained in the evaluated set. For decision support, this distinction matters: an adviser can be reasonably confident that increasing engagement consistency and improving early assessment performance generally lowers academic risk according to the evaluated models, but the magnitude of change required to move a specific student below an institutional alert threshold depends on the model that defines that threshold.

4.8. Controlled Comparison with Established Counterfactual Methods

The evaluation so far establishes internal consistency and the contribution of individual components, but not whether the integrated generator offers an empirical advantage over existing counterfactual methods. To settle that question, the generator was compared against two established methods reproduced from their official implementations: DiCE [24], in its random and genetic variants, and NICE [41], with its sparsity and proximity rewards. A common evaluation protocol was applied across all five methods. The dataset, temporal cutoffs, group-based leakage-controlled split, prediction model, validation-optimized thresholds, target enrolments, actionable feature set, percentile bounds, counterfactual budget, and scoring functions were kept consistent. However, the comparative evaluation applies the corrected constraint ordering, in which the percentile bound is imposed before the directional check rather than after it. The comparison therefore preserves a common evaluation basis while distinguishing the constraint ordering used here from the one used in the main experiments. Two adjustments were required for the comparison to be fair rather than merely uniform. Both libraries decide class membership at a predicted probability of 0.5, whereas this study operates at validation-optimized thresholds of 0.42, 0.48, and 0.53, so a strictly monotone piecewise-linear map was applied that sends each threshold to 0.5; a baseline that moves an instance across the rescaled boundary therefore satisfies exactly the validity condition used throughout this paper, and every reported risk was recomputed from the original, unrescaled model output. The external implementations also emit continuous features at a finite decimal precision. A feature was therefore counted as changed only when the move exceeded both a small fraction of its robust scale and half of the smallest value that feature can represent, so that output rounding is not recorded as a substantive change to a student profile.

The comparative evaluation applies the corrected constraint ordering, in which percentile-based feature bounds are imposed before the directional check rather than after it. The counterfactuals reported in Sections 4.3 to 4.7 were produced under the earlier ordering, among which four directional violations were identified in 1,080 generated counterfactuals. The absence of such violations here therefore does not retroactively amend those results, which are reported as generated. Not every constraint could be imposed equally, and the differences are stated explicitly below. DiCE accepts the actionable feature set and the directional constraints through its feature-mutability setting and a per-instance range constraint, but offers no plausibility filter and no sparsity cap. NICE has no feature-mutability interface at all, so it was run in its published form, and its counterfactuals were then projected onto the actionable subspace, with non-actionable features reset to the student's original value and validity re-evaluated after projection. The effect of that projection is reported with the results.

Table 12 and Figure 7 report the pooled comparison over 360 target enrolments per method. Besides DiCE-random, the proposed generator was the only method to produce a valid counterfactual for every target enrolment at every cutoff; DiCE-genetic reached 0.9056, NICE-proximity 0.8944, and NICE-sparsity 0.7750, with the shortfall concentrated at the earliest and most operationally valuable cutoff, where NICE-sparsity served only 58.3% of students. Because every method is scored on the same target enrolments, coverage is a paired binary outcome rather than two independent proportions. McNemar exact tests on the

discordant enrolments confirm the gap for both NICE variants at every cutoff, with the largest p-value 0.016 and four of the six comparisons below 0.001, and for DiCE-genetic at day 30 and day 90 ($p < 0.001$ and $p = 0.031$), while the day-60 difference for DiCE-genetic was not significant ($p = 0.250$). Paired bootstrap intervals over 20,000 resamples agree with every one of these verdicts. The same pairing applies to validity, measured per enrolment as the share of a student's returned counterfactuals that are valid: the proposed generator reaches 1.0000 at every cutoff, and the paired bootstrap separates it from every baseline except DiCE-random, which also reaches 1.0000, with differences from 0.0514 to 0.4167 and p no larger than 0.0024. In all nine comparisons, each discordant enrolment was one the proposed generator served, and the baseline did not; no baseline recovered recourse that the proposed generator missed. The clearest separation lies in the size of the recourse: mean predicted-risk reduction was 0.5277 for the proposed generator against 0.2178 to 0.2461 for the baselines, a factor of 2.1 over the strongest of them, and the margin widened with the cutoff (0.4385, 0.5467, and 0.5980). Paired Wilcoxon signed-rank tests on the same enrolments place every one of these differences at $p < 0.001$. Joint plausibility was 1.0000 for the proposed generator at all three cutoffs, against 0.9194 for DiCE-random.

Table 12. Controlled comparison of counterfactual generation methods

Method	Cov.	Valid.	Risk red.	Spars.	Robust prox.	Plaus.	Divers.	Time (s)
Proposed	1.0000	1.0000	0.5277	4.7361	1.2925	1.0000	0.9328	2.21
DiCE-random	1.0000	1.0000	0.2424	2.0583	1.6302	0.9194	1.4074	0.70
DiCE-genetic	0.9056	0.9817	0.2178	5.1250	0.4174	0.9989	0.2439	3.46
NICE-sparsity	0.7750	0.8254	0.2461	2.2219	0.7739	0.9793	0.0000	0.07
NICE-proximity	0.8944	0.8994	0.2284	4.2654	0.4453	0.9777	0.0000	0.14

Note: Coverage denotes the proportion of target enrolments receiving at least one valid counterfactual. Plausibility represents the joint range and calibrated nearest-neighbor criteria. Diversity is the mean pairwise robust distance among a student's counterfactuals. Generation time is reported in seconds per target enrolment.

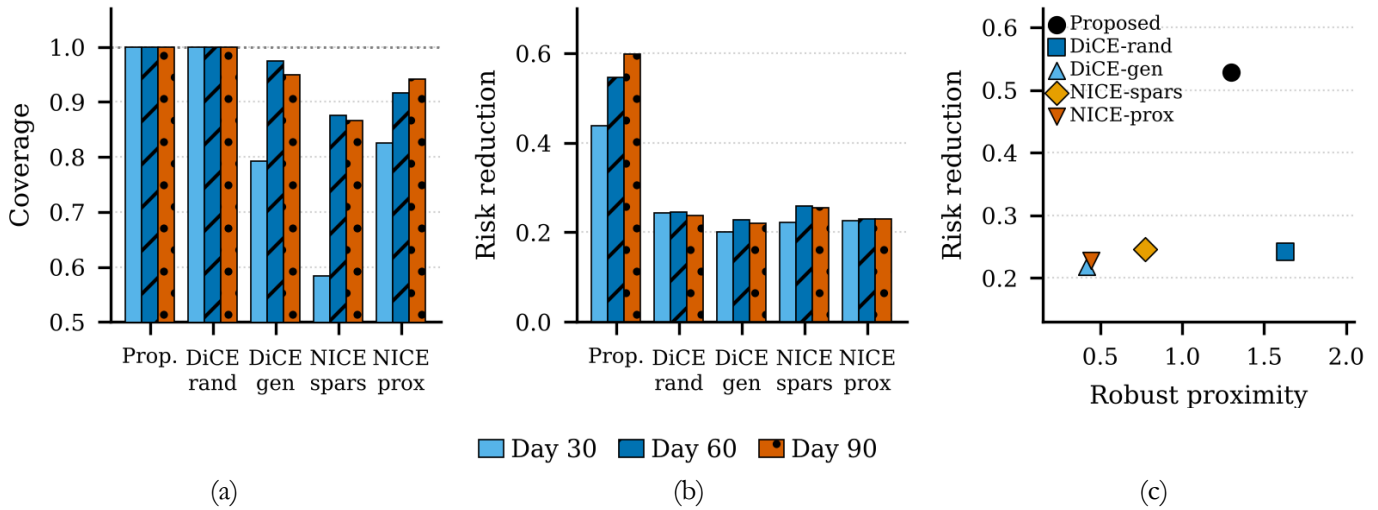

Figure 7. Controlled comparison against reproduced counterfactual baselines under a common OULAD protocol: (**a**) proportion of target enrolments with valid counterfactuals; (**b**) mean predicted-risk reduction; (**c**) trade-off between robust proximity and risk reduction. Computational cost is reported in Table 12.

Part of the risk-reduction gap follows from where each method aims rather than from how well it searches. DiCE and NICE target the decision boundary, which is what a minimal counterfactual is defined to be, whereas the proposed generator anchors each counterfactual to a favorable neighbor that lies inside the favorable region. A larger risk reduction and a costlier robust proximity are therefore two readings of one design choice, and the five methods are better understood as occupying different positions on a shared trade-off curve than as ranked on a single axis. The comparison shows plainly what the framework gives up on that curve. DiCE-genetic produced the most economical counterfactuals, with a robust

proximity of 0.4174 against 1.2925, and NICE-proximity followed closely at 0.4453; on this criterion the proposed generator is the most expensive of the five. Read together with sparsity, the picture is less one-sided, because DiCE-genetic changed more features than the proposed generator, 5.1250 against 4.7361, while moving each of them less, which yields a mathematically closer counterfactual but a more diffuse instruction to a student. Diversity separates the methods further: 0.9328 for the proposed generator against 0.2439 for DiCE-genetic and zero for NICE, which returns a single counterfactual per student by construction, so the proposed generator supplies an adviser with genuinely distinct options where the sparsity-optimizing baselines supply one. Computational cost runs the other way. At 2.21 s per enrolment the proposed generator is roughly thirty times slower than NICE-sparsity at 0.07 s and three times slower than DiCE-random at 0.70 s, although it remains faster than DiCE-genetic at 3.46 s. For cohort-scale deployment this cost, rather than recourse quality, is the binding practical constraint, and it is the price of filtering every candidate profile for plausibility before a counterfactual is retained.

The most consequential result concerns actionability rather than any single metric. Before projection onto the actionable subspace, both NICE variants achieved a validity rate of 1.0000, because a nearest-instance method always locates a real training profile of the opposite class. After projection, validity fell to 0.8254 for the sparsity reward and 0.8994 for the proximity reward, since 22.8% and 14.5% of their counterfactuals respectively changed at least one feature a student cannot change, such as gender, age band, disability, or number of previous attempts. Roughly one in five recommendations from the sparsity-optimizing baseline is therefore unusable as student recourse, not because the method is weak but because it was never designed to respect mutability. The same limitation appears in the directional constraints, which NICE cannot express at all: 15.7% and 48.6% of its counterfactuals move at least one feature in the prohibited direction. This is the clearest evidence that the actionability layer performs substantive work rather than restating a constraint the baselines would satisfy anyway. The same check was applied to the proposed generator. It returns no counterfactual that moves a feature in a prohibited direction or touches a non-actionable feature: zero across the 1,080 counterfactuals of this comparison, and zero across the 1,080 counterfactuals generated under the three base prediction models of Table 13. The directional constraint and the percentile bound can pull against each other, and the generator resolves them in the order that preserves the directional guarantee. The bound is applied first and the direction is rechecked afterwards, in the random-candidate path and in the favorable-neighbor path alike, so a student already beyond the 99th percentile on a feature constrained to increase has that feature left unchanged rather than moved down to the bound; the feasible set for that feature is empty, and the generator selects another. The re-check is inactive for any student whose value already lies inside the percentile band, which holds for all but two of the 360 target enrolments in this comparison. DiCE-random retains a residue of 0.56% of its counterfactuals containing a change in the prohibited direction, and DiCE-genetic none. This check belongs to the corrected constraint ordering used in the comparative and cross-model experiments and does not retroactively amend the four violations identified in the main experiments.

Table 13 addresses whether the generator depends on the predictor that drives it, by running it with Random Forest and Logistic Regression as base models in addition to XGBoost, each with its own validation-optimized threshold and its own selected at-risk enrolments. Joint plausibility was 1.0000 in all nine configurations, so the calibrated plausibility component is independent of the choice of predictor. Coverage and risk reduction are not. Logistic Regression achieved complete coverage with the largest risk reductions, from 0.6028 to 0.6785, but also demanded the largest changes, with a robust proximity of 1.9358 at day 30, whereas Random Forest was consistently the weakest, with coverage of 0.7250, 0.8000, and 0.9750 and risk reductions between 0.2660 and 0.3503. This is the same conservatism identified in Sections 4.7 and 4.9: Random Forest combines the highest false-negative rate with the lowest false-positive rate, and a decision boundary is placed that conservatively requires a larger probability shift before recourse is accepted. The two analyses were conducted independently and agree, which supports reading the behaviour as a property of the estimator rather than of the generator.

Taken together, these results support a bounded claim. Under a common OULAD early-warning protocol, the integrated generator provides complete recourse coverage, calibrated plausibility at every cutoff, and roughly twice the predicted-risk reduction of established

alternatives. These figures come from the gradient-boosted base model used throughout, and of the three properties, only calibrated plausibility transfers unchanged to every base model examined: complete coverage and the risk-reduction margin hold under Logistic Regression but not under Random Forest, where coverage falls to 0.7250 at day 30 and risk reduction to between 0.2660 and 0.3503, close to the baseline range rather than twice it. It does not produce the most economical counterfactuals, and it is not the fastest. The comparison is also confined to one dataset, one prediction task, and the constraint settings described above; DiCE and NICE were designed for different objectives, and their weaker showing on coverage and plausibility reflects a protocol that enforces constraints they do not target rather than any intrinsic deficiency. What the experiment establishes is that those constraints are not free: when actionability, directionality, and calibrated plausibility are enforced on a common footing, the integrated design retains recourse for every student while three of the four reproduced baselines lose between 9% and 23% of them.

Table 13. Generator quality under three base prediction models

Cutoff	Base model	Coverage	Validity	Risk red.	Robust prox.	Plaus.
30	XGBoost	1.0000	1.0000	0.4246	1.4659	1.0000
	Random Forest	0.7250	0.7167	0.2660	0.8558	1.0000
	Logistic Regression	1.0000	1.0000	0.6028	1.9358	1.0000
60	XGBoost	1.0000	1.0000	0.5463	1.2855	1.0000
	Random Forest	0.8000	0.7833	0.2867	0.8009	1.0000
	Logistic Regression	1.0000	1.0000	0.6584	1.8257	1.0000
90	XGBoost	1.0000	1.0000	0.5918	1.1369	1.0000
	Random Forest	0.9750	0.9583	0.3503	0.8288	1.0000
	Logistic Regression	1.0000	1.0000	0.6785	1.3274	1.0000

Note: Each base model uses its own validation-optimized threshold and its own selected at-risk enrolments, with 40 target enrolments per cutoff. Mean sparsity was between 4.1167 and 4.8583 in every configuration.

4.9. Discussion

The results indicate that plausibility-aware counterfactual explanations can translate risk predictions into intervention-oriented decision support. Rather than merely identifying students at risk, the framework suggests actionable directions, including improving assessment performance, increasing consistency of VLE access, broadening resource use, and submitting assignments earlier. The resulting adviser-facing records present the predicted risk, behavioral changes associated with a lower predicted risk, plain-language recommendations, and corresponding support categories. For example, a day-30 student in the high-risk stratum with a predicted risk of 0.81 was associated with a counterfactual that reduced the predicted risk to 0.10 by increasing active VLE days from 4 to 34, distinct resources accessed from 12 to 36, and engagement with course content and the homepage. These changes were translated into recommendations to improve platform access consistency, broaden learning-resource use, and increase content engagement, under the Engagement support category. In the intended human-in-the-loop workflow, the model flags the student, the counterfactual generator proposes plausibility-constrained behavioral targets, and the risk stratum indicates the potential intensity of support. The academic adviser then interprets and prioritizes the recommendations before deciding on appropriate actions. This example illustrates the intended operation of the framework, not a validated improvement in student outcomes, as its practical usefulness has not yet been evaluated with advisers or students. The counterfactual targets are consistent with SHAP-based risk drivers, supporting the internal coherence between explanations and recommendations. This finding is consistent with prior explainable learning analytics studies identifying assessment performance and engagement indicators as central predictors of student success and risk [18]–[20].

The predictive performance, with F1-scores ranging from about 0.69 at day 30 to 0.73 at day 90, is moderate, which is understandable in an early-warning setting. At day 30, limited course activity provides weaker behavioral signals, whereas additional engagement and assessment evidence becomes available by day 90. The inherently imbalanced task and reliance on behavioral rather than administrative signals also constrain achievable F1. Nevertheless,

moderate discrimination does not necessarily undermine the counterfactual component: recourse is generated only for borderline at-risk cases where the model is reasonably calibrated, as reflected in the low Brier scores, and its direction remains stable across independent model families, although the required magnitude varies. Error analysis of baseline-threshold XGBoost predictions further shows that the dominant error mode changes with the cutoff. At day 30, the false-negative rate (0.327) exceeds the false-positive rate (0.244), indicating more missed at-risk students. By day 90, these rates decrease to 0.277 and 0.165, respectively. Model-specific differences are also evident: at day 90, Random Forest has the highest false-negative rate (0.344), despite the lowest false-positive rate (0.116), reflecting a conservative boundary that misses more at-risk students. XGBoost provides a more balanced trade-off, whereas the Dummy baseline never identifies the minority class. This pattern is consistent with the reduced Brier scores and supports later cutoffs when higher precision is required.

The improvement across cutoffs was assessed using 1000-resample bootstrap confidence intervals. The 95% intervals for F1 at day 30 ([0.677, 0.704]) and day 90 ([0.715, 0.746]) do not overlap; the corresponding ROC-AUC intervals are also disjoint, indicating a meaningful improvement in predictive performance. Model differences were further examined using two-proportion z-tests on test-set error counts. XGBoost achieved significantly higher precision than Logistic Regression at every cutoff (for example, day 90: 0.7438 versus 0.6745, $z = 4.91$, $p < 0.001$) and significantly higher recall than Random Forest (day 90: 0.7234 versus 0.6562, $z = 4.64$, $p < 0.001$). Random Forest attained higher precision at the cost of significantly lower recall, reflecting its more conservative decision boundary. These results indicate that XGBoost provides the most favorable recall-precision balance for identifying at-risk students. However, since the models were evaluated on the same test instances, the comparisons are paired, and a McNemar test on per-student agreement would provide a stricter assessment, as in the coverage comparison in Section 4.8. The reported proportion tests should therefore be interpreted as indicating the direction and approximate magnitude of differences rather than exact significance levels.

The controlled comparison suggests that the five methods differ less in their ability to generate counterfactuals than in whether the resulting explanations remain actionable. Before projection onto the actionable subspace, both NICE variants were perfectly valid; the shortfall emerged only after mutability and directional constraints were enforced. Thus, the distinguishing contribution of the integrated design lies primarily in its constraint layer, which determines which features students can realistically influence and in what direction. From an information systems perspective, the framework functions as an explainable analytics layer that transforms predictive risk scores into intervention-oriented support, complementing learning analytics dashboards with adviser-facing intervention categories. This aligns with prior studies advocating a shift from descriptive reporting toward predictive and prescriptive decision support [9]–[11] and supports educational business intelligence by integrating data, predictive modelling, explanation, and action planning [38].

Improved plausibility, however, comes at the cost of higher sparsity. The mean sparsity of approximately 4.7 changed features suggests that realistic recourse may involve multiple academic intervention dimensions rather than single-feature advice. Since changing about five dimensions simultaneously may be difficult for students, counterfactuals are better interpreted as prioritized options rather than simultaneous prescriptions. Advisers can begin with the one or two highest-impact changes, based on the per-feature contribution ordering, and introduce further adjustments incrementally. Importantly, predicted risk reductions do not establish that the suggested behavioral changes will necessarily cause improved academic outcomes. The framework should therefore support, rather than replace, academic advisers, consistent with ethical learning analytics principles emphasizing transparency, student agency, and responsible institutional use of predictive models [35]–[37].

The cross-model analysis in Section 4.7 further qualifies the interpretation of recourse. Isotonic and Platt scaling preserved validity at 1.0000 across all cutoffs, indicating that calibration was not the main weakness. However, recourse strength remains estimator-dependent: a counterfactual that clears the XGBoost alert threshold may not clear that of another model. Consequently, the required intervention magnitude should not be treated as an intrinsic property of the student. The predictor and its alert threshold should therefore be fixed before recourse targets are communicated, and the recommendations require validation under actual academic interventions before their effectiveness can be established.

Several limitations should be considered. First, OULAD represents an online distance-learning context, so the findings may not generalize directly to face-to-face, blended, or local institutional settings. Second, the risk-stratified sample comprised 120 targeted at-risk enrolments per cutoff, representing 119 distinct students, and was selected to make counterfactual search computationally tractable and balance the strata. Thus, the 100% coverage describes this curated pool rather than the full population of predicted at-risk students; scaling recourse generation to the entire cohort remains future work. Third, calibrated nearest-neighbor plausibility depends on the selected 99th-percentile threshold. Although the sensitivity analysis showed that recourse metrics were invariant to this choice and that the percentile mainly controls the strictness of plausibility enforcement, the appropriate level of strictness still requires institutional judgment. Fourth, the intervention categories were derived from feature mappings and have not yet been evaluated by academic advisers or students. Fifth, the cross-model analysis re-scores 357 of the 360 counterfactuals per cutoff. The counterfactual identifier encodes the cutoff, the student, and the counterfactual index but not the module presentation, so at each cutoff the one student who holds two enrolments produces three colliding identifiers; re-scoring retains one of the two enrolments and omits the other three counterfactuals. The generating model reproduced a validity rate of exactly 1.0000 under this reconstruction. The four counterfactuals that move a feature in a prohibited direction each belong to a student with a single enrolment and were therefore among the 357 re-scored cases. Sixth, the controlled comparison in Section 4.8 covers two counterfactual families on a single dataset. Other generators, such as MOC, What-If, and group-counterfactual approaches, were not reproduced; therefore, the reported margins represent evidence under the specified protocol rather than a general ranking of methods. Seventh, the counterfactuals reported in Sections 4.3 to 4.7 were generated using an implementation that applied the percentile bound after the directional check rather than before it. Consequently, four counterfactuals move a feature in the prohibited direction: four changes among 5,098 feature changes, in four of 1,080 counterfactuals and three of the 120 targeted enrolments at day 90. These cases are reported as generated. Recomputing the reported means without those four counterfactuals changes mean predicted-risk reduction at day 90 from 0.6101 to 0.6095, mean sparsity from 4.7222 to 4.7191, and robust proximity from 1.2229 to 1.2231, while coverage, validity, and plausibility remain at 1.0000. Thus, the counterfactual quality results reported in Sections 4.3 to 4.5 do not materially depend on these cases. The cross-model re-scoring in Section 4.7 was not recomputed, as this would require the four counterfactuals to be re-scored under the alternative models; they constitute four of the 357 counterfactuals re-scored per cutoff. Future studies should validate the framework using institutional datasets, stakeholder evaluation, and longitudinal intervention outcomes.

5. Conclusions

This study developed an explainable counterfactual decision-support framework for early academic risk prediction that connects XGBoost-based prediction, SHAP-based interpretation, actionability constraints, calibrated nearest-neighbor plausibility, robust proximity evaluation, and personalized intervention mapping, positioned as a human-in-the-loop decision-support component rather than an autonomous intervention system. Predictive performance improved as more learning evidence became available, with validation-optimized XGBoost that improved in both discrimination and calibration at later cutoffs (full metrics reported in Section 4). The generated recommendations satisfied the validity and plausibility criteria across all evaluated strata, although four directional-constraint violations occurred in the initial implementation at day 90. The results also indicate that later cutoffs generally enabled greater predicted-risk reduction without a substantial increase in robust proximity. A cross-model analysis showed that the direction of academic recourse is robust across model families whereas the magnitude sufficient to cross an institutional alert threshold is estimator-dependent, and a controlled comparison against reproduced DiCE and NICE baselines under a common protocol found that the framework retained recourse for every targeted student, at calibrated plausibility and roughly twice their predicted-risk reduction, while remaining the slower and less economical option. The research impact is to shift explainable learning analytics from explaining why a student is at risk toward identifying what actionable and plausible support may reduce predicted risk, offering institutions an explainable analytics layer for advising dashboards; because the counterfactuals are model-based rather than causal guarantees,

future work should validate the framework using local institutional datasets, adviser and student evaluation, and longitudinal outcomes on retention and academic success.

Author Contributions: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing—original draft preparation, and Visualization: J.B.; Validation: J.B., G.J.M., T.M.T. and S.; Writing—review and editing: G.J.M., T.M.T. and S.; Supervision: J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: This study used the publicly available Open University Learning Analytics Dataset (OULAD) [33], distributed under a Creative Commons Attribution 4.0 licence and obtainable from https://analyse.kmi.open.ac.uk/open_dataset. No new data were created. All experiments were run with a fixed random seed (42) and a group-based, leakage-controlled temporal split, so the data preparation, the splits, and the counterfactual generation procedure are reproducible from the dataset and the configuration reported in Section 3. Exact numerical reproduction carries one caveat: the fitted gradient-boosted and logistic models depend on the numerical libraries beneath them, and the validation-optimized threshold is the quantity most exposed to this, because the F1 curve is broad near its maximum. At day 60, for instance, every threshold between 0.43 and 0.50 lies within 0.002 of the maximum F1, so refitting in a different library environment can move the selected threshold while leaving performance materially unchanged. The thresholds reported in Section 4.1 are those used for every counterfactual analysis in this paper and should be treated as fixed inputs when reproducing it, rather than re-optimized.

Acknowledgments: The authors acknowledge the Open University Learning Analytics Dataset (OULAD) providers for making the dataset publicly available for research purposes. In the interest of transparency, the authors disclose the use of a generative AI tool (Anthropic Claude) as an assistive aid during manuscript preparation, specifically for language editing, improving the clarity and organization of the text, and reviewing the manuscript for internal consistency. The AI tool was not used to generate, manipulate, or fabricate any research data, to design or execute the experiments, to perform the underlying analyses, or to make scientific decisions. All datasets, experimental procedures, results, and findings originate entirely from the authors' own work, and the authors take full responsibility for the accuracy, content, and integrity of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] C. Romero and S. Ventura, "Educational Data Mining and Learning Analytics: An Updated Survey," *WIREs Data Min. Knowl. Discov.*, vol. 10, no. 3, p. e1355, May 2020, doi: 10.1002/widm.1355.

[2] A. Namoun and A. Alshanqiti, "Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review," *Appl. Sci.*, vol. 11, no. 1, p. 237, Dec. 2020, doi: 10.3390/app11010237.

[3] F. Ouyang and P. Jiao, "Artificial Intelligence in Education: The Three Paradigms," *Comput. Educ. Artif. Intell.*, vol. 2, p. 100020, 2021, doi: 10.1016/j.caeai.2021.100020.

[4] E. Alyahyan and D. Düştegör, "Predicting Academic Success in Higher Education: Literature Review and Best Practices," *Int. J. Educ. Technol. High. Educ.*, vol. 17, no. 1, p. 3, Dec. 2020, doi: 10.1186/s41239-020-0177-7.

[5] B. Albreiki, N. Zaki, and H. Alashwal, "A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques," *Educ. Sci.*, vol. 11, no. 9, p. 552, Sep. 2021, doi: 10.3390/educsci11090552.

[6] J. Kabathova and M. Drlik, "Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques," *Appl. Sci.*, vol. 11, no. 7, p. 3130, Apr. 2021, doi: 10.3390/app11073130.

[7] Y. Liu, S. Fan, S. Xu, A. Sajjanhar, S. Yeom, and Y. Wei, "Predicting Student Performance Using Clickstream Data and Machine Learning," *Educ. Sci.*, vol. 13, no. 1, p. 17, Dec. 2022, doi: 10.3390/educsci13010017.

[8] B. Alnasyan, M. Basheri, and M. Alassafi, "The Power of Deep Learning Techniques for Predicting Student Performance in Virtual Learning Environments: A Systematic Literature Review," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100231, Jun. 2024, doi: 10.1016/j.caeai.2024.100231.

[9] F. Gutiérrez, K. Seipp, X. Ochoa, K. Chiluiza, T. De Laet, and K. Verbert, "LADA: A Learning Analytics Dashboard for Academic Advising," *Comput. Human Behav.*, vol. 107, p. 105826, Jun. 2020, doi: 10.1016/j.chb.2018.12.004.

[10] G. Ramaswami, T. Susnjak, A. Mathrani, and R. Umer, "Use of Predictive Analytics within Learning Analytics Dashboards: A Review of Case Studies," *Technol. Knowl. Learn.*, vol. 28, no. 3, pp. 959–980, Sep. 2023, doi: 10.1007/s10758-022-09613-x.

[11] T. Susnjak, G. S. Ramaswami, and A. Mathrani, "Learning Analytics Dashboard: a Tool for Providing Actionable Insights to Learners," *Int. J. Educ. Technol. High. Educ.*, vol. 19, no. 1, p. 12, Dec. 2022, doi: 10.1186/s41239-021-00313-7.

[12] R. Alamri and B. Alharbi, "Explainable Student Performance Prediction Models: A Systematic Review," *IEEE Access*, vol. 9, pp. 33132–33143, 2021, doi: 10.1109/ACCESS.2021.3061368.

[13] R. Farrow, "The Possibilities and Limits of Explicable Artificial Intelligence (XAI) in Education: a Socio-Technical Perspective," *Learn. Media Technol.*, vol. 48, no. 2, pp. 266–279, Apr. 2023, doi: 10.1080/17439884.2023.2185630.

[14] M. Saqr and S. López-Pernas, "Why Explainable AI May Not Be Enough: Predictions and Mispredictions in Decision Making in Education," *Smart Learn. Environ.*, vol. 11, no. 1, p. 52, Nov. 2024, doi: 10.1186/s40561-024-00343-4.

[15] H. Khosravi *et al.*, "Explainable Artificial Intelligence in education," *Comput. Educ. Artif. Intell.*, vol. 3, p. 100074, 2022, doi: 10.1016/j.caeai.2022.100074.

[16] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Nov. 2017, pp. 4768–4777, [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295230.

[17] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[18] B. Alnasyan, M. Basheri, M. Alassafi, and K. Alnasyan, "Kanformer: An Attention-Enhanced Deep Learning Model for Predicting Student Performance in Virtual Learning Environments," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, p. 25, Mar. 2025, doi: 10.1007/s13278-025-01446-7.

[19] E. Kalita, H. El Aouifi, A. Kukkar, S. Hussain, T. Ali, and S. Gaftandzhieva, "LSTM-SHAP Based Academic Performance Prediction for Disabled Learners in Virtual Learning Environments: a Statistical Analysis Approach," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, p. 65, Jun. 2025, doi: 10.1007/s13278-025-01484-1.

[20] B. Ujkani, D. Minkovska, and N. Hinov, "Course Success Prediction and Early Identification of At-Risk Students Using Explainable Artificial Intelligence," *Electronics*, vol. 13, no. 21, p. 4157, Oct. 2024, doi: 10.3390/electronics13214157.

[21] B. I. Igoche, O. Matthew, P. Bednar, and A. Gegov, "Integrating Structural Causal Model Ontologies with LIME for Fair Machine Learning Explanations in Educational Admissions," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 65–85, Jun. 2024, doi: 10.62411/jcta.10501.

[22] G. T. Ayem, O. Asilkan, and A. Iorliam, "Design and Validation of Structural Causal Model: A Focus on EGRA Dataset," *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 86–103, Nov. 2023, doi: 10.33633/jcta.v1i2.9304.

[23] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera, "Model-Agnostic Counterfactual Explanations For Consequential Decisions," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Jun. 2020, vol. 108, pp. 895–905.

[24] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Jan. 2020, pp. 607–617, doi: 10.1145/3351095.3372850.

[25] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications," *Inf. Fusion*, vol. 81, pp. 59–83, May 2022, doi: 10.1016/j.inffus.2021.11.003.

[26] S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, and C. Shah, "Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review," *ACM Comput. Surv.*, vol. 56, no. 12, pp. 1–42, Dec. 2024, doi: 10.1145/3677119.

[27] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *SSRN Electron. J.*, 2017, doi: 10.2139/ssrn.3063289.

[28] A. Van Looveren and J. Klaise, "Interpretable Counterfactual Explanations Guided by Prototypes," in *Machine Learning and Knowledge Discovery in Databases. Research Track*, N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, and J. A. Lozano, Eds. Cham: Springer International Publishing, 2021, pp. 650–665.

[29] P. Buñay-Guisñan, J. A. Lara, and C. Romero, "Counterfactual Explanations in Education: A Systematic Review," *WIREs Data Min. Knowl. Discov.*, vol. 16, no. 1, p. e70060, Mar. 2026, doi: 10.1002/widm.70060.

[30] P. Buñay-Guisñan, A. Cano, A. Anguera, J. A. Lara, and C. Romero, "Group Counterfactual Explanations: A Use Case to Support Students at Risk of Dropping Out in Online Education," *Electronics*, vol. 15, no. 1, p. 51, Dec. 2025, doi: 10.3390/electronics15010051.

[31] M. Çavuş and J. Kuzilek, "An Effect Analysis of the Balancing Techniques on the Counterfactual Explanations of Student Success Prediction Models," *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Derg.*, vol. 15, no. Special Issue, pp. 302–317, Nov. 2024, doi: 10.21031/epod.1526704.

[32] F. Afrin, M. Hamilton, and C. Thevathyan, "Exploring Counterfactual Explanations for Predicting Student Success," in *Computational Science -- ICCS 2023*, M. Jiří and de Mulatier, Eds. Cham: Springer Nature Switzerland, 2023, pp. 413–420.

[33] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics Dataset," *Sci. Data*, vol. 4, no. 1, p. 170171, Nov. 2017, doi: 10.1038/sdata.2017.171.

[34] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[35] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Mon. Weather Rev.*, vol. 78, no. 1, pp. 1–3, Jan. 1950, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[36] P. Prinsloo and S. Slade, "Student Vulnerability, Agency and Learning Analytics: An Exploration," *J. Learn. Anal.*, vol. 3, no. 1, pp. 159–182, Apr. 2016, doi: 10.18608/jla.2016.31.10.

[37] A. Pardo and G. Siemens, "Ethical and Privacy Principles for Learning Analytics," *Br. J. Educ. Technol.*, vol. 45, no. 3, pp. 438–450, May 2014, doi: 10.1111/bjet.12152.

[38] S. Gaftandzhieva, S. Hussain, S. Hilcenko, R. Doneva, and K. Boykova, "Data-driven Decision Making in Higher Education Institutions: State-of-play," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 6, 2023, doi: 10.14569/IJACSA.2023.0140642.
[39] N. Hidayat, L. Afuan, and H. R. Jannah, "Prescriptive Learning Analytics for Student Dropout: Integrating Temporal Velocity and Counterfactual Explanations in Longitudinal Data," *J. Comput. Theor. Appl.*, vol. 3, no. 4, pp. 627–646, May 2026, doi: 10.62411/jcta.15920.
[40] D. R. I. M. Setiadi, A. R. Muslikh, S. W. Iriananda, W. Warto, J. Gondohanindijo, and A. A. Ojugo, "Outlier Detection Using Gaussian Mixture Model Clustering to Optimize XGBoost for Credit Approval Prediction," *J. Comput. Theor. Appl.*, vol. 2, no. 2, pp. 244–255, Nov. 2024, doi: 10.62411/jcta.11638.
[41] D. Brughmans, P. Leyman, and D. Martens, "NICE: an Algorithm for Nearest Instance Counterfactual Explanations," *Data Min. Knowl. Discov.*, vol. 38, no. 5, pp. 2665–2703, Sep. 2024, doi: 10.1007/s10618-023-00930-y.