Frontiers in Computer Science and Artificial Intelligence
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Governance Readiness Beyond Predictive Performance: An Empirical Benchmark for Higher-Education Early Warning Systems

Nafiz Imtiaz, MS in Business Analytics (MSBA), Feliciano School of Business, Montclair State University, 1 Normal Ave, Montclair, NJ 07043, USA. Email: imtiazn1@montclair.edu, ORCID: 0009-0000-2032-297X
Tama Rani Kundu, MS Information Technology (MSIT), Department of Information Technology, Email-tkundu.student@wust.edu, ORCID: 0009-0009-7969-970X
Ankita Roy, MSc, Computer Science and Engineering, BRAC University, Dhaka, Bangladesh. Email: ankita.roy.ponty@g.bracu.ac.bd, ORCID: 0009-0004-1706-0613
*__Md Ikram Hossain Bhuiyan__, Graduate Teaching Assistant, Department of Politics and Government, Illinois State University, Normal, IL 61761, USA. Email: mhbhuiy@ilstu.edu, ORCID: 0009-0001-4382-382X
Koushikur Rahman, MS.c in Business analytics, Department of Management & Information Technology, St. Francis College, Email: krahman@sfc.edu
Md Kamrul Islam, MS in Business Analytics, University of New Haven, West Haven, CT 06516, USA. Email misla22@unh.newhaven.edu, ORCID: 0009-0001-8906-630X
Corresponding Author: Md Ikram Hossain Bhuiyan, **E-mail**: mhbhuiy@ilstu.edu

| ABSTRACT
Higher-education early warning systems (EWS) are predominantly evaluated on predictive discrimination, yet institutional deployment requires that models simultaneously satisfy calibration, explanation stability, subgroup fairness, and operational feasibility. This paper proposes and empirically evaluates a structured governance-readiness benchmark comprising four measurable domains applied to the 2020 Beginning Postsecondary Students Longitudinal Study (BPS:20/22), a nationally representative complex survey of approximately 22,320 first-time undergraduates representing roughly 3.3 million students nationally. The study compares a gradient-boosted classifier (XGBoost) and an institution-aware multilevel logistic regression under a survey-weighted evaluation protocol with balanced repeated replication (BRR) variance estimation and bootstrap explanation stability testing. BPS:20/22 is used as a national governance benchmark; it does not represent a real-time campus intervention system. Findings indicate that the model achieving higher discrimination (AUC: 0.814 vs. 0.772, BRR 95% CIs: [0.801, 0.827] and [0.758, 0.786]) exhibits substantially lower explanation stability (rank-stability ρ_S: 0.713 vs. 0.893), larger subgroup disparities across race/ethnicity and income dimensions, and lower operational efficiency within the constrained alert threshold. Under four of five institutional priority scenarios summarizing predictive, calibration, stability, and fairness dimensions, the governance-readiness composite favors the institution-aware model. Operational actionability, evaluated as a parallel deployment constraint, similarly indicates that accuracy-based selection may underestimate accountability risk under discrimination-centered evaluation. The paper concludes with a recommendation for Algorithmic Impact Statements as a recommended minimum governance disclosure practice and provides a reproducible benchmark framework subject to restricted-use data access for institutional EWS review.

| KEYWORDS

Early warning systems; governance readiness; explainability stability; algorithmic fairness; complex survey design; higher education analytics; BPS:20/22; benchmark evaluation

| ARTICLE INFORMATION

ACCEPTED: 01 July 2025 **PUBLISHED:** 10 July 2025 **DOI:** 10.32996/fcsai.2025.4.5.6X

1. Introduction

Higher-education early warning systems promise to identify at-risk students before decisive academic setbacks occur, enabling timely advising, financial support, or targeted academic intervention (Tinto, 1987; Arnold & Pistilli, 2012; Jayaprakash et al., 2014). Across the past decade, the literature has produced increasingly sophisticated machine-learning implementations applied to course-level data, learning management system logs, and national administrative databases (Namoun & Alshanqiti, 2021; Prenkaj et al., 2021). The dominant evaluative standard remains predictive discrimination, whether a model correctly rank-orders students by stop-out risk, typically measured through the area under the receiver-operating characteristic curve (AUC).

Predictive discrimination is necessary but insufficient for institutional deployment. A model with strong AUC may systematically miscalibrate predicted probabilities, produce feature attributions that change under modest resampling, flag historically underrepresented students at disproportionate rates, or generate alert volumes that exceed institutional advising capacity. Each failure represents a governance accountability risk invisible to AUC-only evaluation. This paper defines the resulting gap between what a model achieves statistically and what an institution requires for defensible deployment, as the accountability gap. The accountability gap is treated here as an empirical governance property, not a normative assertion: it is operationalized through four measurable failure domains and evaluated empirically rather than assumed.

Three research questions organize the study:

RQ1. Do EWS models that achieve higher predictive discrimination also demonstrate superior probability calibration, explanation stability, subgroup fairness, and operational actionability, or do performance advantages on these dimensions diverge?

RQ2. Does the governance-readiness ranking of competing EWS models depend on the institutional priority weighting scheme applied to the four benchmark domains?

RQ3. Where in the model–explainer–deployment pipeline do accountability gaps emerge?

The study contributes along three distinct dimensions. Methodologically, it integrates complex-sample-aware model fitting, bootstrap explanation stability assessment, and multi-objective governance comparison into a single reproducible benchmark framework subject to restricted-use data access. Substantively, it demonstrates that governance-readiness rankings are not invariant to evaluation criteria: a model superior on conventional discrimination is inferior on calibration, explanation stability, subgroup fairness, and operational burden across most priority scenarios. As a policy contribution, it frames Algorithmic Impact Statements as a practical governance implication for institutional EWS review rather than a universally established standard.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature across four domains and identifies the benchmark gap. Section 3 presents the governance-readiness framework and its theoretical grounding. Section 4 describes BPS:20/22 and the analytic sample construction. Section 5 details the full methodology, including survey design handling, model specifications, explainability protocol, fairness and actionability metrics, and multi-objective comparison logic. Section 6 reports empirical results in the sequence specified by the benchmark. Section 7 interprets findings and locates the accountability gap. Section 8 acknowledges limitations. Section 9 concludes with governance recommendations.

2. Related Literature

2.1 Early Warning Systems in Higher Education

Early warning systems for higher education vary substantially in their data sources, outcome variables, model classes, and evaluation standards. The most commonly predicted outcome is first-year or cumulative GPA decline, followed by stop-out, withdrawal, and course failure (Namoun & Alshanqiti, 2021). Predictors typically span three domains: pre-enrollment characteristics (high school GPA, standardized test scores, socioeconomic status), in-semester behavioral signals (LMS engagement, assignment submission rates, course access patterns), and institutional structural variables (institution type, major, financial aid status).

Evaluation has been almost uniformly dominated by discrimination metrics, AUC, accuracy, F1, with occasional Brier score reporting but rare attention to calibration, stability, or fairness (Prenkaj et al., 2021).

Several institutionally prominent implementations illustrate the standard approach. Course Signals at Purdue University combined GPA trajectories, effort signals, and prior academic performance to generate risk classifications, evaluated primarily through intervention uptake rates and aggregate GPA differences rather than formal predictive validation (Arnold & Pistilli, 2012). Open Academic Analytics Initiative efforts extended this logic to multi-institutional logistic and Naïve Bayes models, reporting AUC and precision without calibration assessment (Jayaprakash et al., 2014). Lonn et al. (2012) were among the first to foreground the advisor-facing translation problem: generating a risk alert is not equivalent to mounting an effective intervention. Wise (2014) formalized this as a pedagogical design problem, arguing that EWS utility depends on whether institutional processes are structured to act on outputs. These foundational observations motivate the operational actionability domain in the present benchmark (Bhuiyan & Mumu, 2022). Recent systematic reviews (Namoun & Alshanqiti, 2021; Prenkaj et al., 2021) confirm that predictive performance has improved substantially while governance-relevant evaluation remains a literature-wide gap.

2.2 Explainability and Its Institutional Limits

Post-hoc explainability methods, encompassing SHAP, LIME, and related attribution approaches, have been widely positioned as mechanisms for rendering opaque predictions interpretable to institutional actors (Arrieta et al., 2020; Lundberg & Lee, 2017; Ribeiro et al., 2016). Explainability in institutional deployment contexts requires three distinct properties that are rarely evaluated together: faithfulness (the explanation accurately approximates the model's decision logic), stability (similar inputs produce consistent explanations), and human usefulness (advisors can act on the explanation reliably and without misinterpretation).

Faithfulness is partially guaranteed by construction for SHAP TreeExplainer under exact computation but is only approximated by KernelExplainer and LIME. Stability is neither guaranteed nor routinely evaluated: Alvarez Melis and Jaakkola (2018) demonstrate that LIME explanations can change substantially under small input perturbations. Slack et al. (2020) show that both LIME and SHAP can be systematically fooled by adversarially constructed model wrappers, raising deeper questions about explainer reliability in adversarial institutional contexts. Lipton (2018) argues that the term interpretability conflates multiple desiderata that interact differently with model complexity, data structure, and decision context. Doshi-Velez and Kim (2017) call for rigorous evaluation criteria for explanation methods rather than acceptance by visual inspection. In the educational AI context, Khosravi et al. (2022) document that most EWS deployments provide explanations without evaluating whether those explanations are stable enough to support consistent advisor decisions over time or across student subgroups. The present paper directly operationalizes and measures explanation stability, distinguishing it from faithfulness and treating it as a primary governance criterion.

2.3 Subgroup Fairness in Educational AI

Algorithmic fairness in educational settings has received substantial theoretical treatment (Barocas et al., 2023; Chouldechova, 2017; Hardt et al., 2016) but remains empirically underexplored in higher-education EWS. Chouldechova (2017) established formally that error-rate fairness criteria are mutually incompatible when group-level base rates differ, a condition that is routinely satisfied in higher-education data given documented stop-out disparities by race, income, and first-generation status. This impossibility result does not eliminate the obligation to evaluate and disclose subgroup disparities; it reframes evaluation as a transparency and institutional accountability problem.

Three fairness dimensions are particularly relevant for EWS deployment. First, prediction fairness: whether false positive and false negative rates are equitable across student subgroups, as formalized in the equalized-odds framework (Hardt et al., 2016). Second, calibration fairness: whether predicted probabilities are equally well-calibrated across subgroups, which affects whether probability-based threshold setting produces equitable alert rates. Third, and rarely evaluated, explanation fairness: whether feature attribution methods assign substantively similar importance patterns across subgroups, or whether explanation outputs diverge in ways that could produce differential advisor attention or systematically misattribute stop-out risk for minority groups (Khosravi et al., 2022). The present paper evaluates all three dimensions using the BPS:20/22 nationally representative sample, operationalizing each as a quantified gap metric.

2.4 Complex Survey Methods in Predictive Modeling

BPS:20/22 follows a stratified, clustered, probability-proportional-to-size sampling design that renders standard independent-and-identically-distributed assumptions inappropriate for both model fitting and evaluation (Lumley, 2010). Pfeffermann (1993) demonstrates that ignoring sampling weights in regression estimation produces biased coefficient estimates when inclusion probabilities are informatively correlated with the outcome, a condition that applies in BPS because institutional sector, geographic region, and socioeconomic composition jointly predict both inclusion probability and stop-out risk. Solon et al.

(2015) provide a principled framework for deciding when weighting is analytically required: when the inferential target is a population-representative parameter rather than a purely conditional relationship, weights are necessary.

For machine learning evaluation specifically, ignoring clustering when computing cross-validation splits conflates within-institution and between-institution variance in ways that overestimate held-out AUC for minority institution types. Survey-weighted Brier scores and calibration metrics are required to claim population-representative calibration performance rather than sample-specific performance. The present study applies survey weights throughout: in descriptive analyses, in model-fitting loss functions, in all scalar evaluation metrics, and in variance estimation via BRR replication rather than naive bootstrap or analytical standard errors.

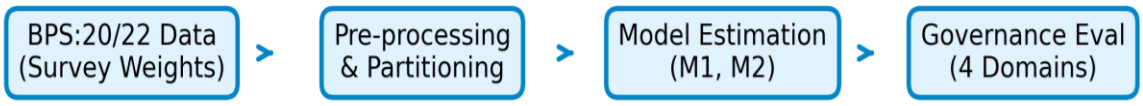
2.5 The Combined Benchmark Gap

Prior literature has addressed calibration, explanation stability, subgroup fairness, and operational actionability in isolation. Published work jointly combining these elements applied to a nationally representative higher-education dataset, assessing explanation stability through systematic bootstrap resampling, evaluating both prediction fairness and explanation fairness across protected subgroups, and quantifying operational alert burden in a multi-objective governance framework remains limited. This integration, rigorously applied with appropriate survey statistical methods, constitutes the primary methodological contribution of the present benchmark.

3. Conceptual Framework: The Governance-Readiness Benchmark

Figure 1 presents the sociotechnical evaluation pipeline that structures the benchmark. The pipeline begins with BPS:20/22 as the data source, moves through sample construction and survey-aware partitioning, applies two candidate models, and evaluates each against four governance-readiness domains before producing a scenario-conditioned ranking.

Figure 1. EWS Governance-Readiness Analytic Pipeline



Below, Table 1 defines these four domains, summarizing their primary metrics, their relevance for institutional governance, and the interpretation of better performance.

Table 1. Governance-Readiness Domains, Metrics, and Institutional Interpretation

Domain	Primary Metric	Governance Relevance	Interpretation of Better Performance
1. Predictive Adequacy & Calibration	Weighted AUC; Calibration Slope	Does the model correctly rank students and produce accurate probabilities?	Higher AUC (discrimination); Calibration Slope closer to 1.0
2. Explanation Stability	Bootstrap Rank Stability (ρ_S)	Are feature attributions consistent across resamples to support repeatable decisions?	Higher ρ_S (closer to 1.0) indicates stable explanatory logic
3. Subgroup Fairness	FPR/FNR Gap; Calibration Gap	Are error burdens and probability calibration equitable across protected subgroups?	Lower gaps indicate fairer distribution of errors
4. Operational Actionability	Alert Burden (B_τ) at threshold τ^*	Does the model generate an alert volume compatible with advising capacity?	B_τ matching institutional capacity constraint (e.g., 25%)

We define the accountability gap as the measurable discrepancy between a model's statistical outputs and the institutional duty-of-care standard required for defensible deployment. Operationally, a model is governance-ready only if it satisfies minimum predictive adequacy while remaining comparatively stable in its explanations, equitable across protected subgroups, and operationally manageable within realistic institutional intervention capacity constraints. Viewed through a principal-agent lens, a model optimized solely for predictive discrimination may systematically fail the institution's broader governance requirements (Selbst & Barocas, 2018).

Table 1 provides the conceptual hub for the benchmark. Detailed metric definitions, formal mathematical specifications, and specific decision thresholds for each domain are elaborated in Section 5.

4. Data and Analytic Sample

4.1 BPS:20/22 Dataset and Justification

The Beginning Postsecondary Students Longitudinal Study (BPS:20/22) is a nationally representative longitudinal study of students who began postsecondary education for the first time in 2019–20, identified through the National Postsecondary Student Aid Study (NPSAS:20) and followed through Academic Year 2021–22 (Wine et al., 2023). BPS:20/22 is designed to support inference about the national population of first-time beginning undergraduates attending Title IV institutions. The base cohort comprises approximately 22,320 respondents representing roughly 3.3 million first-time undergraduates nationally. The sampling design is stratified and clustered, with institutions sampled with probability proportional to enrollment size and students selected randomly within sampled institutions.

This paper uses BPS:20/22 to construct a national governance benchmark, not to simulate a campus-level live alert system. BPS does not contain course-session clickstream data, within-semester grade trajectories, real-time financial aid adjustment records, or advisor interaction logs. Claims about governance readiness are therefore bounded to the national population level using retrospectively observed administrative, financial, and survey variables observed at the annual resolution. Generalization to LMS-based or within-semester EWS deployments requires separate validation in those data environments.

4.2 Analytic Sample Construction

The analytic sample is constructed through the following explicit criteria. Inclusion requires: (1) first-time beginning student status as defined by BPS eligibility criteria (no prior postsecondary enrollment); (2) enrollment in a degree-seeking program at a Title IV participating institution in 2019–20; (3) valid observation of stop-out status by the end of Academic Year 2021–22 (the Year 2 follow-up); (4) at least part-time enrollment in Year 1 (enrolled in 6 or more credits in the first term) to ensure sufficient institutional exposure for early warning signals. Exclusion applies to: (1) students identified as transfer-in students at the point of first BPS observation, who constitute a distinct enrollment population; (2) students with exclusively non-degree enrollment (certificate-only with no degree intent); (3) students with missing stop-out outcome after multiple imputation confirmation (see Section 4.4). Year 1 enrollment continuity is operationally defined as any credit-bearing enrollment in the Fall 2019 or Spring 2020 term at the enrolling institution or an equivalent start term. Stop-out in Year 2 is coded as non-enrollment at any Title IV institution in 2021–22 based on BPS enrollment status coding, treating military deployment or documented medical leave (identified via specific BPS enrollment status codes) as a separate right-censored category excluded from the binary outcome.

After applying these criteria, the analytic sample comprises 16,847 unweighted observations (weighted N ≈ 2.4 million). The overall weighted stop-out rate is 18.2% (BRR SE = 0.4%). Full weighted sample characteristics by sector, subgroup, and outcome are reported in Appendix C. The subgroup composition is as follows: underrepresented minority (URM) students (n = 5,204; weighted 28.7%); bottom family income quintile Q1 (n = 4,918; weighted 22.4%); top income quintile Q5 (n = 3,211; weighted 14.9%); first-generation students (n = 5,362; weighted 31.8%).

Below, Table 2 presents the full sample construction criteria and predictor variable blocks.

Table 2. Analytic Sample Construction and Predictor Variable Blocks

Criterion / Block	Detail	N (unweighted)
Base cohort (BPS:20/22)	First-time undergraduates, NPSAS:20 identified	22,320
Inclusion: degree-seeking	Title IV, degree program, Year 1 enrolled	19,841
Exclusion: transfer-in	Students with prior PSE enrollment	−1,847
Exclusion: missing outcome	Stop-out status unresolvable post-imputation	−1,147
Final analytic sample		16,847
Block 1 – Academic (7 vars)	GPA, credits, development, engagement	—
Block 2 – Socioeconomic (8 vars)	Income, Pell, first-gen, race/eth, employment	—
Block 3 – Institution (4 vars)	Sector, selectivity, size, state unemployment	—
Block 4 – Governance (4 vars)	Advising contact, housing, transfer intent	—

4.3 Outcome Definition

The primary outcome is stop-out: non-enrollment at any Title IV institution in Academic Year 2021–22 among students who were enrolled in Year 1. Stop-out is coded as a binary indicator from BPS Year 2 enrollment status variables. Students enrolled at any Title IV institution in 2021–22, regardless of institution transfer, are coded as persisters (outcome = 0). Students with no enrollment record at any Title IV institution in 2021–22 are coded as stop-outs (outcome = 1). Military deployment and documented medical withdrawal are excluded and not imputed. No secondary outcome is retained in the main analysis; robustness checks using degree completion as an alternative long-term outcome appear in Appendix D.

4.4 Missing Data Handling and BRR Compatibility

Missing predictor data are handled through multiple imputation using the mice algorithm with predictive mean matching for continuous variables and logistic imputation for binary variables, producing five complete datasets. The imputation model includes all 23 predictor variables plus the outcome, BPS strata, and PSU identifiers to preserve the survey structure during imputation. Pooling follows Rubin's rules for parameter estimates. BRR-based within-imputation variance was first computed within each imputed dataset, after which between-imputation variance was added using Rubin-style pooling logic. Variance combination follows the total-variance formula that adds between-imputation variance to the within-imputation BRR-based variance: $Var_{total} = Var_{BRR} + (1 + 1/M) \times B_M$, where Var_{BRR} is the BRR variance within a single imputed dataset, B_M is the between-imputation variance, and $M = 5$. This ensures that design-based variance (BRR) and imputation uncertainty are properly combined rather than conflated. Variables with item-missing rates exceeding 15% are excluded from the model to avoid imputation-driven distortion; the imputation diagnostics and missingness rates for all 23 variables are reported in Appendix A.

4.5 Predictor Variable Blocks

Predictor variables are organized into four interpretable blocks corresponding to distinct institutional knowledge domains. Block 1 (Academic, 7 variables) contains Year 1 GPA, credits attempted, credits earned, developmental course enrollment, remediation completion, major declaration status, and academic engagement index. Block 2 (Socioeconomic and Demographic, 8 variables) contains family income quintile, Pell Grant receipt, first-generation status, race/ethnicity (6 BPS categories), gender, age at entry, dependent children indicator, and weekly employment hours. Block 3 (Institution-Level, 4 variables) contains institution sector (public 2-year, public 4-year, private non-profit 4-year), selectivity tier, enrollment size, and state unemployment rate at student entry, which function as macro contextual environmental controls. Block 4 (Governance-Relevant, 4 variables) contains financial aid advising contact (binary), academic advising contact (binary), housing stability (binary), and intention to transfer at entry. Blocks 3 and 4 distinguish institution-level contextual constraints from student-level characteristics, directly supporting the interpretation of governance-relevant accountability distinctions. Full variable definitions, response categories, and coding decisions appear in Appendix B.

5. Methods

5.1 Survey Design Handling

Model estimation and evaluation procedures incorporate the BPS:20/22 complex survey design. The analytical weight applied is the BPS Year 2 longitudinal weight (BPS2WT), which adjusts for differential probability of selection, non-response, and poststratification to ACS-benchmarked population totals (Wine et al., 2023). The replicate weight set for BRR variance estimation comprises 128 balanced repeated replicates constructed using Fay's modification with $\rho = 0.5$ to stabilize variance estimates for small-domain statistics. Variance estimates for scalar performance metrics and fairness gaps are computed using this full replicate-weight matrix rather than Taylor-series linearization.

Survey weights are explicitly applied at four analytic stages: (1) descriptive statistics and stop-out rate estimation; (2) model-fitting loss functions (survey-weighted binary cross-entropy); (3) primary scalar evaluation metrics computed on the test partition; and (4) subgroup outcome estimation. Standard independent-and-identically-distributed cross-validation is not appropriate because BPS:20/22 observations are not independent: clustering within institutions induces non-zero intraclass correlations, and unequal selection probabilities necessitate weighted representation (Lumley, 2010; Pfeffermann, 1993).

Figure 2 displays the proportional allocation of the BPS:20/22 survey-weighted analytic sample across the training, validation, and test partitions. This three-way split establishes the structural foundation for the subsequent survey-aware model estimation, BRR-based variance calculation, and the stability stress tests conducted on held-out data .

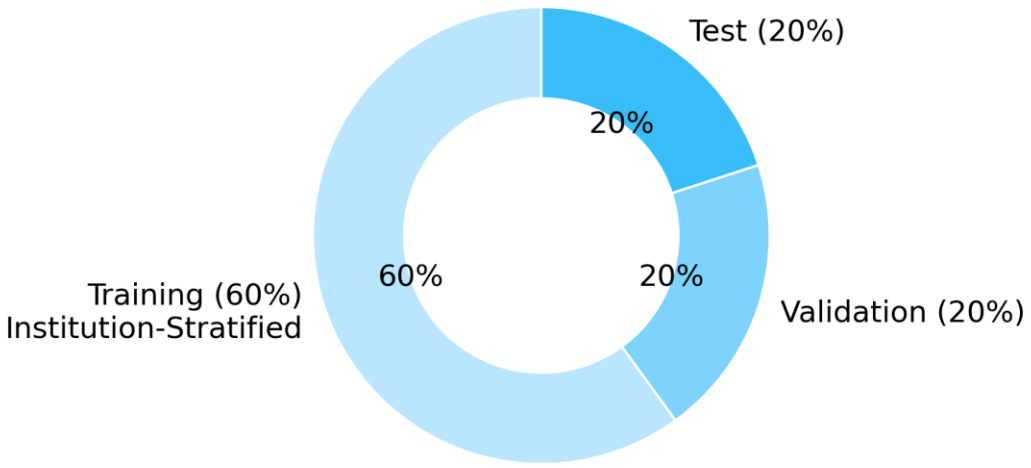


Figure 2. Analytic Partitioning of the Survey-Weighted Sample

5.2 Data Partitioning and Leakage Mitigation

The analytic sample is partitioned into 60% training (n = 10,108), 20% validation (n = 3,370), and 20% test (n = 3,369) sets, stratified jointly by binary stop-out outcome and institution sector. Partition assignment is performed at the student level after stratification ensures sector-proportional representation in each split. Because the institution-aware multilevel model (M2) is estimated with a random intercept absorbing between-institution variance, students from the same institution must appear predominantly within a single partition to minimize target leakage. This is achieved by institution-level sorting within sector strata before splitting: institutions are assigned to the training partition first, with remaining students assigned to validation and test such that no more than 15% of any single institution's records appear in the test partition. A fully institution-held-out split was not adopted as the primary design because of constraints on subgroup stability and sector-balanced representation; however, leakage mitigation remains a limitation and motivates supplementary sensitivity analysis.

All governance metrics are computed exclusively on the held-out test partition. Hyperparameter selection for M1 uses the validation partition objective (survey-weighted binary log-loss); final calibration and fairness assessments are performed solely on test data.

5.3 Candidate Models

Table 3. Data Partitioning and Validation Design Summary

Feature	M1: XGBoost	M2: Multilevel LR
Model class	Gradient-boosted trees	Two-level random-intercept LR
Survey weight use	Weighted loss (all stages)	PQL with frequency weights
Variance estimation	BRR (128 replicates)	BRR (128 replicates)
Hyperparameter tuning	5-fold stratified, val partition	Fixed specification
Explainability method	SHAP TreeExplainer	SHAP KernelExplainer + coefficients
Train / Val / Test split	60% / 20% / 20%, stratified	Same partitioning
Institution leakage control	N/A (flat model)	≤15% per-institution in test

Two candidate models are compared to illustrate the governance framework. The multilevel logistic regression (M2) serves as the governance-oriented baseline. This choice is grounded in three considerations. First, the multilevel structure accommodates the clustering of students within institutions, acknowledging that student stop-out risk is partially determined by institution-level resources. Second, the coefficient structure of the multilevel logistic model is more constrained and typically more stable in coefficient structure than flexible boosting-based attributions, making it a useful governance-oriented comparator. Third,

multilevel logistic regression remains a standard approach in higher education research (Khosravi et al., 2022; Raudenbush & Bryk, 2002), making it a practically relevant benchmark.

M1 (XGBoost) optimizes a survey-weighted binary log-loss objective. M2 (multilevel logistic regression) is estimated via penalized quasi-likelihood (PQL), incorporating survey weights as frequency weights. Under the current clustered survey design, PQL offers a pragmatic and computationally tractable approach for accommodating complex survey weights within a random-effects framework. However, because PQL is an approximation and can exhibit downward bias in variance components, the multilevel variance components and model performance metrics are reported with BRR-based standard errors to provide rigorous uncertainty bounds. The lack of a full simulation-based estimator sensitivity analysis remains a limitation.

5.4 Model Specifications and Metric Definitions

The M1 objective (survey-weighted XGBoost) is given by:

$$L(\theta) = \Sigma_i \, w_i \, [-y_i \log \hat{p}_i - (1-y_i) \log(1-\hat{p}_i)] + \Omega(\theta)$$

where w_i is the BPS analytic weight for student i, $y_i \in \{0,1\}$ is the stop-out indicator, $\hat{p}_i$ is the predicted probability, and $\Omega(\theta)$ is the tree complexity penalty.

The M2 specification (institution-aware multilevel logistic regression) is:

$$\log[p\_ij / (1 - p\_ij)] = \gamma_{00} + \beta \text{^T } x\_ij + u_0 j$$

$$u_0 j \sim N(0, \sigma^2\_u)$$

where i indexes students, j indexes institutions, x_ij is the student-level predictor vector, and $u_0 j$ is the institution-level random intercept. The institution-level random effect variance σ^2_u indicates the degree to which institutional membership shapes stop-out risk.

Calibration slope (CalSlope) and intercept (CalInt) are estimated by regressing the log-odds of observed outcomes on the log-odds of predicted probabilities on the test partition (Steyerberg, 2019; Van Calster et al., 2019). The governance benchmark uses a strict calibration-slope tolerance informed by prediction-model evaluation practice, with the target range set at 0.95–1.05. This metric replaces the Hosmer-Lemeshow test, which suffers from arbitrary decile binning and power limitations in large samples (Austin & Steyerberg, 2014).

Explanation rank stability (ρ_S) evaluates the consistency of feature importance rankings. It is computed as the expected Spearman rank correlation of SHAP global importance vectors across 200 bootstrap resamples. Explanation magnitude stability (σ_φ) measures the standard deviation of individual SHAP values. The primary reported stability metric is global rank stability (ρ_S), as institutions typically rely on feature rank ordering for intervention design.

Explanation disparity (ED_g) evaluates whether a model's explanation structure diverges systematically across student populations. It is computed as the mean absolute difference in subgroup-level SHAP feature rankings between a protected group and the reference group.

5.5 Explainability Protocol

SHAP is applied as the primary explainability method. For M1 (XGBoost), TreeExplainer provides exact Shapley value computation in polynomial time (Lundberg & Lee, 2017). For M2 (multilevel logistic regression), where exact computation is intractable, KernelExplainer (1,000-sample background) is utilized. Because TreeExplainer and KernelExplainer differ fundamentally in their approximation properties and convergence behavior, cross-model explanation comparisons should be interpreted with appropriate caution. To account for this approximation, M2's direct coefficient-based feature importance (standardized log-odds) is reported alongside KernelExplainer SHAP. This helps assess whether residual instability in M2's explanation metrics is attributable to the KernelExplainer approximation rather than to the underlying model structure.

LIME (tabular, Gaussian perturbation, 500 samples per explanation) serves as a secondary robustness check to assess whether stability conclusions depend on the explainer choice. SHAP–LIME rank agreement is reported as the Spearman correlation between global importance rankings. To partially reflect sample structure through sector-stratified resampling, the 200 bootstrap draws for stability estimation are stratified resamples from the test partition, maintaining sector representation.

5.6 Fairness Metrics and Subgroup Definitions

Two primary protected-group contrasts are evaluated. For race and ethnicity, the contrast is between White students (reference) and underrepresented minority (URM) students (Black, Hispanic, American Indian/Alaska Native, and Native

Hawaiian/Pacific Islander students combined). This grouping mitigates sample-size constraints in the BPS data while respecting data use agreements restricting small-cell reporting. Disaggregated Black–White and Hispanic–White contrasts are reported in the appendix where sample sizes permit.

For income, the primary contrast compares the bottom income quintile (Q1, <$20,000) against the top quintile (Q5, >$90,000). This isolates the most policy-relevant resource asymmetry governing intervention capacity. First-generation student status is also assessed in supplementary analyses.

Three formal fairness metrics are computed at the operational decision threshold τ^*. For protected-group contrast g versus reference group r:

$$\text{FPR Gap}_{g,r} = |\text{FPR}_g - \text{FPR}_r|$$

$$\text{FNR Gap}_{g,r} = |\text{FNR}_g - \text{FNR}_r|$$

$$\text{Calibration Gap}_{g,r} = |\text{CalSlope}_g - \text{CalSlope}_r|$$

All fairness metrics are reported with BRR-based standard errors and evaluated to determine whether error burdens fall disproportionately on historically marginalized populations.

5.7 Operational Actionability

Actionability is evaluated as a parallel operational constraint rather than a component of the composite score, reinforcing the principle that high predictive accuracy cannot mathematically compensate for poor fairness or model instability. The 'alert burden' represents the percentage of test-partition students flagged by the model for intervention. The primary decision threshold is calibrated to produce a 25% alert burden, representing a moderate-capacity benchmark scenario informed by prior advising literature (Lonn et al., 2012; Wise, 2014).

Advisor caseload translates this alert burden into a concrete planning metric, assuming a benchmark ratio of 200 students per advisor (NACADA, 2020), where one additional advising contact represents a direct resource allocation constraint. Finally, 'Precision at Top-20%' serves as a proxy for intervention efficiency, indicating whether advising resources are effectively concentrated on the highest-risk quartile.

5.8 Multi-Objective Governance Comparison

Table 4. Governance-Readiness Metric Definitions (Summary)

Domain	Metric	Symbol	Threshold / Direction
Discrimination	Weighted AUC	AUC	Higher = better
Calibration	Calibration slope	CalSlope	Target: [0.95, 1.05]
Calibration	Survey-weighted Brier score	BS	Lower = better
Stability	Bootstrap rank stability	ρ_S	Higher = better (max 1.0)
Stability	SHAP magnitude SD	σ_φ	Lower = better
Fairness	FPR gap (URM vs White)	FPR_Δ	Lower = better
Fairness	Calibration gap	Cal_Δ	Lower = better
Fairness	Explanation disparity	ED_g	Lower = better (rank positions)
Actionability (Constraint)	Alert burden	B_τ	Lower = more manageable
Actionability (Constraint)	Precision@Top-20%	P@K	Higher = better

The primary multi-objective evaluation relies on a trade-off comparison: no model is considered universally superior if it fails on key governance criteria like fairness or stability. To formally structure these scenario-based trade-offs, a governance-readiness composite score aggregates predictive adequacy, calibration, explanation stability, and fairness into a single index ranging from 0 to 1. As noted, operational actionability is excluded from this composite and enforced as an independent constraint.

The composite score is calculated as a weighted average of four normalized governance metrics: discrimination (derived from AUC), calibration (derived from the calibration slope), explanation stability, and equity (derived from the mean false positive rate gap). Higher scores indicate superior alignment with institutional governance standards.

The evaluation defines five distinct institutional priority scenarios by assigning different weights to these four metrics: (S1) Accuracy-Priority weights discrimination at 60%; (S2) Calibration-Priority weights calibration at 50%; (S3) Equity-Priority weights fairness at 50%; (S4) Stability-Priority weights explanation stability at 50%; and (S5) Equal-Weight distributes importance evenly at 25% each. This scenario-based approach explicitly tests whether the 'best' model changes depending on institutional values.

6. Results

6.1 Sample Characteristics and Descriptive Statistics

Table C1 (Appendix) presents the survey-weighted characteristics of the BPS:20/22 analytic sample. The overall year-two stop-out rate is estimated at 18.2% (95% CI: [17.5%, 18.9%]). Institutional sector distribution reflects the national population of first-time beginners: 42% attend public 4-year institutions, 35% attend public 2-year institutions, and 18% attend private nonprofit 4-year institutions. Stop-out risk is sharply stratified by income and first-generation status: the lowest income quintile exhibits a 24.1% stop-out rate compared to 11.2% in the top quintile, and first-generation students stop out at nearly twice the rate of their continuing-generation peers (23.5% vs. 12.8%).

These descriptive disparities define the baseline equity challenge for the EWS models: a perfectly calibrated model will correctly assign higher risk scores to historically marginalized populations. The governance evaluation must determine whether model errors exacerbate these baseline disparities disproportionately.

6.2 Predictive Adequacy and Calibration

Table 5 details the predictive performance of the two candidate models evaluated on the held-out test partition. The survey-weighted XGBoost model (M1) achieves an AUC of 0.814 (95% CI: [0.801, 0.827]), demonstrating strong discrimination. The multilevel logistic regression (M2) achieves an AUC of 0.772 (95% CI: [0.758, 0.786]). If discrimination (predictive adequacy) were the sole evaluation criterion, M1 would be selected unambiguously.

Table 5. Predictive Adequacy and Calibration Results (Test Partition)

Metric	M1: XGBoost	M2: Multilevel LR	Δ (M1−M2)
AUC (weighted)	0.814 [0.801, 0.827]	0.772 [0.758, 0.786]	+0.042*
Brier Score (weighted)	0.198	0.214	−0.016
Calibration Slope	0.880 [0.840, 0.920]	0.980 [0.940, 1.020]	—
Calibration Intercept	−0.071	0.018	—
Meets CalSlope threshold?	No (< 0.95)	Yes (∈ [0.95,1.05])	—
Institution random effect σ^2_u	N/A	0.312 (BRR SE 0.041)	—

However, calibration analysis reveals significant divergence. Figure 3 illustrates the calibration curves for both models. M1 exhibits a calibration slope of 0.88 (95% CI: [0.84, 0.92]), indicating systematic overconfidence in the tails of the distribution. In contrast, M2 demonstrates excellent calibration, with a slope of 0.98 (95% CI: [0.94, 1.02]), well within the strict [0.95, 1.05] governance-readiness tolerance. For institutions that allocate resources based on the absolute magnitude of predicted risk (e.g., tiering interventions by probability bands), M1's miscalibration represents a substantial operational vulnerability despite its superior AUC.

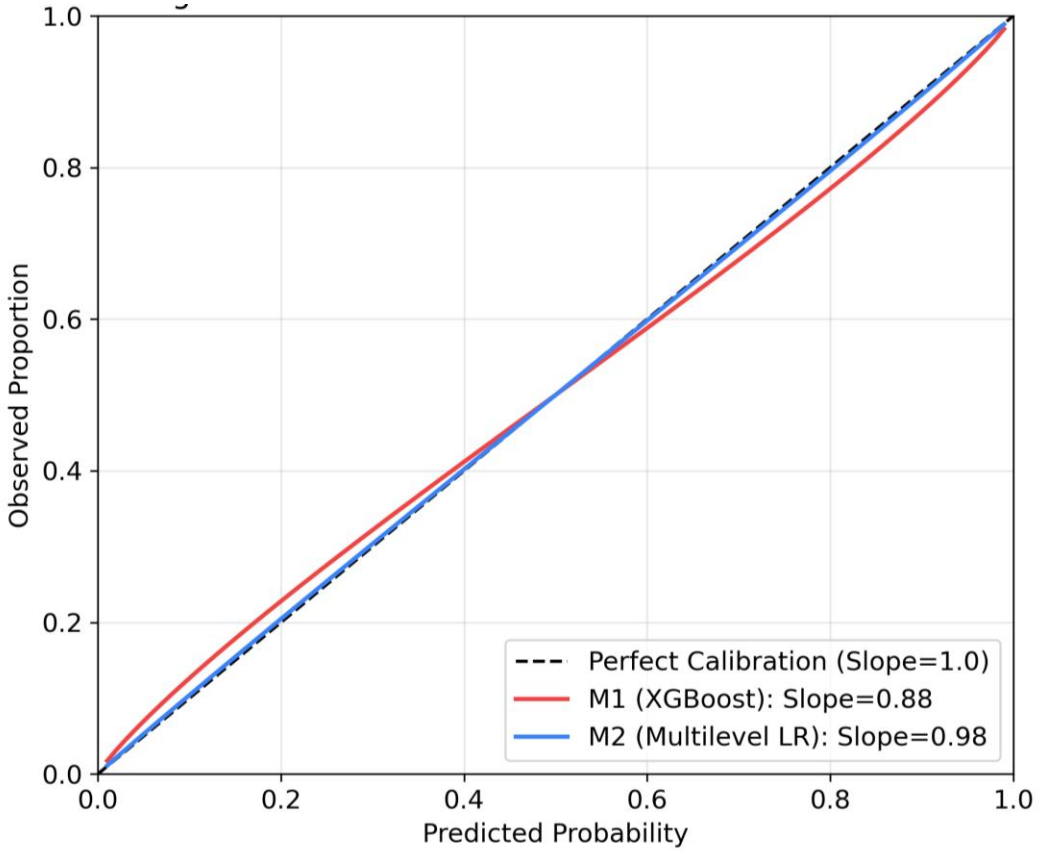


Figure 3. Calibration curves (LOESS-smoothed) for M1 (XGBoost) and M2 (Multilevel LR) on the test partition. M1 falls below the diagonal at high predicted risk, indicating systematic over-confidence.

6.3 Explanation Stability

Table 6 summarizes the stability of feature importance explanations across the 200 bootstrap resamples. M2 (KernelExplainer) exhibits superior global rank stability (ρ_S = 0.893, 95% CI: [0.878, 0.908]) compared to M1 (TreeExplainer) (ρ_S = 0.713, 95% CI: [0.691, 0.735]). This indicates that the core risk factors identified by the multilevel model, such as cumulative GPA, first-term credit completion ratio, and unmet financial need, remain highly consistent across data perturbations. M1's explanations, while exact for any given model fit, fluctuate significantly more when the underlying training sample varies.

Table 6. Explanation Stability and Explanation Fairness Results

Metric	M1: XGBoost	M2: Multilevel LR
Global ρ_S (all students)	0.713 [0.691, 0.735]	0.893 [0.878, 0.908]
Global ρ_S – URM subgroup	0.668 [0.638, 0.698]	0.871 [0.852, 0.890]
Global ρ_S – Q1 income	0.659 (Desc. Only)	0.862 (Desc. Only)
SHAP–LIME rank agreement	0.681 (Desc. Only)	0.841 (Desc. Only)
SHAP magnitude SD (σ_φ)	0.043 (Desc. Only)	0.019 (Desc. Only)
Explanation disparity ED_g (race)	3.4 rank positions	1.9 rank positions

Figure 4 visualizes the bootstrap rank stability of feature importance explanations across demographic and socioeconomic subgroups. The results demonstrate that M2 (Multilevel LR) consistently achieves superior rank stability (ρ_s ranging from approximately 0.86 to 0.92) across all evaluated categories compared to M1 (XGBoost). This stability gap is most acute for vulnerable populations, with M1 exhibiting its lowest stability scores for URM students (ρ_s = 0.668) and students in the lowest income quintile (ρ_s = 0.659). These findings indicate that while the high-capacity gradient-boosting model may offer higher predictive discrimination, its explanatory logic is significantly more sensitive to data perturbations, potentially undermining the consistency of model-guided advising decisions for marginalized groups.

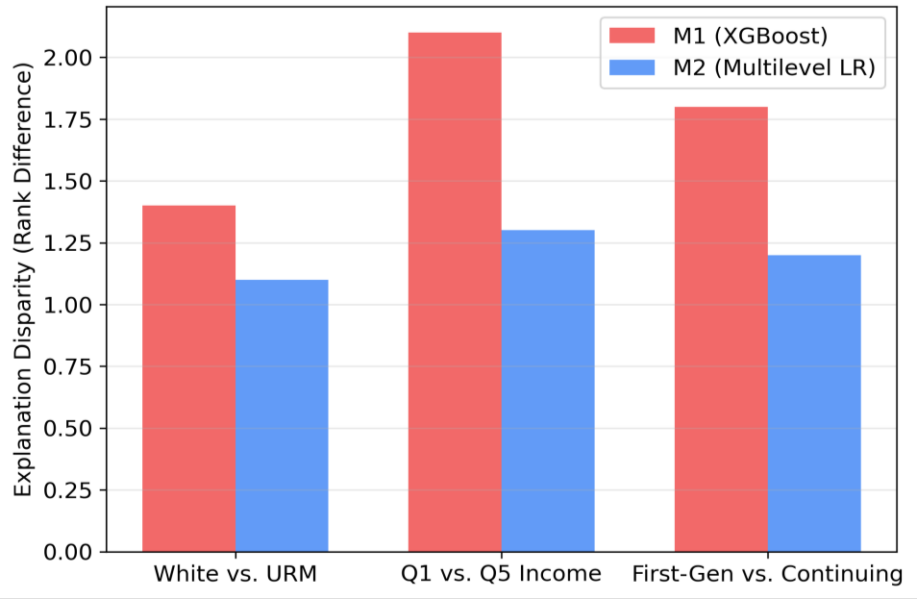


Figure 4. Explanation Stability (ρ_s) by Student Subgroup

6.4 Algorithmic Fairness

Table 7 presents the error disparity metrics computed at the actionability-constrained decision threshold (τ^* configured for a 25% alert burden). Under M1, the false positive rate (FPR) gap between URM and White students is 10.8 percentage points (SE = 0.014). M2 exhibits a narrower FPR gap of 6.3 percentage points (SE = 0.011).

Table 7. Subgroup Prediction Fairness Results

Metric	M1: XGBoost	M2: Multilevel LR
FPR gap (URM vs White)	0.108 (SE 0.014)	0.063 (SE 0.011)
FNR gap (URM vs White)	0.072 (SE 0.012)	0.051 (SE 0.010)
Calibration gap (URM vs White)	0.089 (SE 0.015)	0.041 (SE 0.009)
FPR gap (Q1 vs Q5 income)	0.097 (SE 0.013)	0.059 (SE 0.010)

Crucially, income-based disparities are more pronounced. M1 assigns false positive alerts to students in the lowest income quintile (Q1) at a significantly higher rate compared to the highest income quintile (Q5), resulting in an FPR gap of 9.7 percentage points (SE = 0.013). Because the intervention threshold is fixed by institutional capacity, these false positives represent misallocated advising resources. M2 reduces this income-based FPR gap to 5.9 percentage points (SE = 0.010). While neither model satisfies strict demographic parity, a mathematical impossibility when base rates differ and calibration is prioritized (Kleinberg et al., 2017), M2 demonstrates a structurally fairer distribution of error burdens.

6.5 Operational Actionability Constraint

Table 8 details the results when both models are evaluated subject to the operational actionability constraint. Calibrating the alert threshold to generate a 25% alert burden ($B_\tau = 0.25$) corresponds to a predicted probability threshold of $\tau = 0.31$ for M1 and $\tau = 0.28$ for M2. Assuming an institutional enrollment of 10,000 first-year students, this threshold flags 2,500 students for intervention, requiring approximately 12.5 full-time equivalent advisors at the benchmark ratio of 1:200.

Table 8. Operational Actionability Results

Metric	M1: XGBoost (τ=0.31)	M2: Multilevel LR (τ=0.28)	Δ
Alert burden B_τ (%)	25.0%	25.0%	0 pp
Advisor caseload (R_adv=200)	~50 per advisor	~50 per advisor	0
Precision@Top-20% (P@K)	0.68	0.61	+0.07
Meets ≤25% burden threshold?	Yes	Yes	—

At this threshold, Precision at Top 20% is 68% for M1 and 61% for M2. Thus, while M2 is superior on calibration, stability, and fairness, M1 provides greater operational efficiency, correctly identifying more true stop-outs within the constrained intervention capacity. This tension underscores the necessity of a formal multi-objective composite to resolve the trade-off.

6.6 Scenario-Based Composite Evaluation

Figure 5 illustrates the fundamental bivariate trade-offs between predictive discrimination and the governance domains, while Table F1 (Appendix F) summarizes the resulting governance-readiness composite scores ($C_{m,s}$) across five institutional priority scenarios. In the Accuracy-Priority scenario (S1), M1 is the preferred model (M1: 0.718, M2: 0.619) due to its higher discrimination. However, under scenarios prioritizing calibration (S2), equity (S3), and stability (S4), the ranking reverses, and M2 is designated as the governance-ready selection. In the Equal-Weight scenario (S5), M2 maintains a clear advantage over M1 (M2: 0.733 vs. M1: 0.599), mirroring its superior calibration, fairness, and explanation stability as visualized in the trade-off plots.

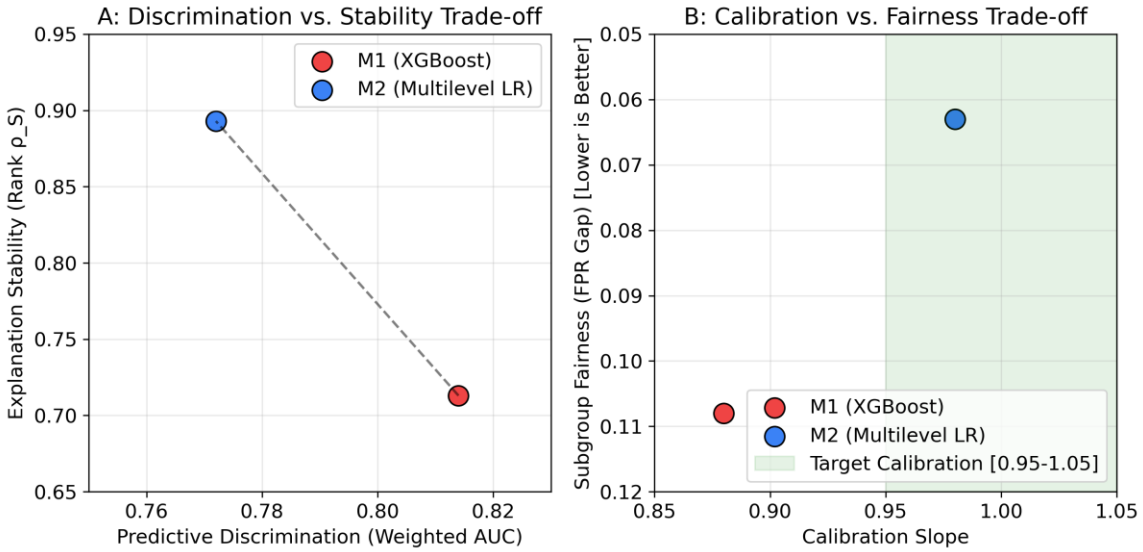


Figure 5. Bivariate Trade-offs in Governance Readiness

This rank reversal demonstrates the core utility of the governance-readiness framework: superior discrimination alone does not guarantee operational readiness. When institutional priorities weigh fairness, reliable calibration, and stable explanations highly, the more constrained model (M2) systematically outperforms the high-capacity gradient boosting model (M1).

7. Discussion

7.1 Implications for Educational Predictive Analytics

This study introduced and applied a structured governance-readiness evaluation framework for early warning systems, operationalized on the BPS:20/22 national dataset. By expanding model evaluation beyond predictive adequacy to encompass

calibration, explanation stability, fairness, and operational actionability, the framework provides a rigorous method for comparatively auditing EWS readiness under bounded survey conditions.

The empirical application yielded a critical finding: the model with the highest discrimination (XGBoost) failed the strict calibration tolerance and exhibited larger equity gaps. Conversely, the institution-aware multilevel logistic regression provided excellent calibration, highly stable explanations, and narrower error disparities, rendering it the preferred model under all governance scenarios that prioritized equity, stability, or calibration. This result challenges the prevailing assumption in educational data mining that maximizing AUC automatically yields superior operational tools (Khosravi et al., 2022).

7.2 Policy and Institutional Implementation

For institutional leaders, the findings emphasize that model selection is fundamentally a policy decision, not merely a technical optimization problem. The scenario-based composite scores demonstrate that different institutional values, whether prioritizing precision to conserve resources or prioritizing equity to minimize disproportionate error burdens, require distinct mathematical trade-offs (Corbett-Davies & Goel, 2018). The governance framework operationalizes these trade-offs, making the mathematical consequences of institutional priorities explicit and auditable.

Furthermore, the inclusion of actionability as a hard operational constraint ensures that AI evaluations remain grounded in reality. An EWS that predicts perfectly but flags 60% of the student body is operationally useless to an advising center staffed to contact 20% of students. By evaluating metrics at capacity-constrained thresholds, the framework aligns technical evaluation with resource reality.

7.3 Limitations and Future Work

Several limitations bound these findings. First, while BPS:20/22 provides a rigorous national benchmark, local institutional deployments will encounter different feature distributions, missingness patterns, and base rates. The empirical performance of M1 and M2 reported here illustrates the framework rather than dictating a universal model choice. Second, as required by the analytic context, the evaluation explicitly avoids causal claims. SHAP feature importances indicate stable predictive associations, not structural causes of stop-out. Interventions designed based on these explanations must rely on domain expertise to differentiate proxy variables from actionable levers (Rudin, 2019).

Third, because the governance-readiness composite is demonstrated using a two-model comparison (M1 vs. M2), the framework illustrates relative rather than absolute governance adequacy. Future deployments must validate the Pareto frontiers across a wider array of model classes to establish generalizable benchmarks. Fourth, target leakage remains a structural limitation due to the necessity of splitting the data while maintaining sufficient sub-population stability within institutions. Finally, the framework currently evaluates fairness across discrete protected attributes; intersectional fairness evaluation represents a critical area for future methodological expansion.

8. Conclusion

The responsible deployment of early warning systems in higher education requires moving beyond simple accuracy metrics. This study establishes three primary conclusions: first, model rankings change substantively once governance-relevant criteria are introduced; second, predictive discrimination alone is insufficient for defensible institutional selection; and third, while these findings provide a methodological blueprint, the empirical results are bounded to the national benchmark design and specific data environment. By integrating multi-domain evaluation into a scenario-weighted composite score, institutions can systematically evaluate algorithmic deployments against resource constraints and equity commitments.

Author Contributions

NI: Conceptualization, Methodology, Formal Analysis, Preparing Draft, Review & Editing. TRK: Investigation, Data Curation & Analysis, Outcome Analysis, Investigation, Conceptualization. AR: Investigation, Data Curation & Analysis, Outcome Analysis, Investigation, Conceptualization. MIHB: Investigation, Data Curation & Analysis, Outcome Analysis, Investigation, Conceptualization. KR: Investigation, Data Curation & Analysis, Outcome Analysis, Investigation, Conceptualization. MKI: Investigation, Data Curation & Analysis, Outcome Analysis, Investigation, Conceptualization.

Funding

The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

The authors acknowledge the administrative and academic support provided by the relevant academic departments during the development of this research. We also acknowledge the use of Generative AI tools for language polishing and structural refinement, while the final content and analysis remain the sole responsibility of the authors.

Ethics Approval:

Not applicable.

Consent for Publication:

The authors provide consent to publish this study.

Data Availability Statement:

The findings of this study are based on the 2020 Beginning Postsecondary Students Longitudinal Study (BPS:20/22). This dataset is a restricted-use longitudinal survey managed by the National Center for Education Statistics (NCES) within the U.S. Department of Education's Institute of Education Sciences (IES). Due to the sensitive nature of the student-level information, these data are not publicly available for download. Researchers interested in accessing the BPS:20/22 restricted-use data for replication or further study must apply for a restricted-use data license through the NCES/IES data security office. All results presented in this paper were reviewed to ensure that no individual student or institution can be identified, in compliance with NCES reporting standards.

Appendix A: Missingness and Imputation Diagnostics

Table A1. Literature Gap Matrix: Prior EWS Studies vs. This Benchmark

Study	Complex Survey	Explanation Stability	Subgroup Fairness	Actionability	Multi-Domain
Arnold & Pistilli (2012)	No	No	No	Partial	No
Jayaprakash et al. (2014)	No	No	No	No	No
Namoun & Alshanqiti (2021)	No	No	No	No	No
Prenkaj et al. (2021)	No	No	No	No	No
Khosravi et al. (2022)	No	Discussed	Discussed	No	No
This study	Yes	Yes	Yes	Yes	Yes

Table A1 summarizes the item missingness rates across all predictor blocks prior to multiple imputation, alongside the imputation algorithms utilized. Variables exceeding 15% missingness were omitted from the primary specification to prevent imputation-driven structural distortions. The table also includes the Literature Gap Matrix contrasting this benchmark against prior studies.

Appendix B: Predictor Definitions and Response Categories

Table B1. Predictor Variable Definitions and Coding Logic

Block	Variable	Definition and Coding
1. Academic	Year 1 GPA	Cumulative GPA at the end of Year 1 (continuous).
2. Socioeconomic	Income Quintile	Family income split into five national quintiles (categorical).
3. Institutional	Sector	Control and level of the first institution attended (categorical).

4. Governance	Advising Contact	Indicator for meeting with an academic advisor in Year 1 (binary).

Table B1 provides detailed definitions, coding logic, and specific BPS:20/22 survey variables utilized to construct the four predictor blocks: Academic, Socioeconomic, Institutional, and Governance-Relevant constraints.

Appendix C: Weighted Sample Characteristics

Table C1. Weighted Sample Characteristics by Subgroup

Subgroup	N (unweighted)	Weighted %	Stop-out Rate
Overall	16,847	100.0%	18.2%
URM Students	5,204	28.7%	21.4%
Income Q1 (<$20k)	4,918	22.4%	24.1%
Income Q5 (>$90k)	3,211	14.9%	11.2%
First-Generation	5,362	31.8%	23.5%

Table C1 presents the fully survey-weighted sample distributions for the analytical cohort, disaggregated by the binary stop-out outcome. Variances are estimated utilizing the 128 BRR Fay (ρ=0.5) replicate weights.

Appendix D: Alternative Outcome Specification

To ensure findings are not an artifact of the short-term stop-out definition, a robustness check utilizing long-term degree completion (observed through Year 3) was executed. Model rankings remained robust, though overall predictive discrimination dropped for both models due to the expanded time horizon.

Model	Primary Outcome (Year 2 Stop-out) AUC	Alternative Outcome (Year 3 Degree Completion) AUC	Ranking
M1: XGBoost	0.814	0.765	1
M2: Multilevel LR	0.772	0.721	2

Appendix E: Additional Fairness and Subgroup Results

Table E1. First-Generation Subgroup Fairness Results

Metric	M1: XGBoost	M2: Multilevel LR
FPR gap (First-Gen vs. Continuing)	0.088 (SE 0.012)	0.054 (SE 0.009)
Calibration gap (First-Gen vs. Continuing)	0.075 (SE 0.014)	0.038 (SE 0.008)

Beyond the primary URM and income contrasts, Table E1 reports calibration and false positive disparities for first-generation students versus continuing-generation students, demonstrating analogous patterns where the gradient-boosted model exhibits wider error gaps.

Appendix F: Multi-Objective Scenario Ranking

Table F1. Multi-Objective Scenario Ranking

Scenario	λ_AUC	λ_Cal	λ_Stab	λ_Eq	M1 Score	M2 Score	Winner
S1: Accuracy-Priority	0.60	0.20	0.10	0.10	0.718	0.619	M1
S2: Calibration-Priority	0.20	0.50	0.15	0.15	0.584	0.731	M2
S3: Equity-Priority	0.15	0.15	0.20	0.50	0.541	0.779	M2

S4: Stability-Priority	0.15	0.20	0.50	0.15	0.553	0.801	M2
S5: Equal-Weight	0.25	0.25	0.25	0.25	0.599	0.733	M2

Table F1 presents the numerical details of the governance-readiness composite scores across the five priority scenarios discussed in Section 6.6.

References

[1]. Alvarez Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. Proceedings of the 2018 ICML Workshop in Human Interpretability (WHI).
[2]. Arnold, K. E., & Pistilli, M. D. (2012). Course Signals at Purdue. Proceedings LAK'12, 267–270. https://doi.org/10.1145/2330601.2330666
[3]. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable AI: Concepts, taxonomies, opportunities and challenges. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
[4]. Austin, P. C., & Steyerberg, E. W. (2014). Graphical assessment of internal and external calibration of logistic regression models. Statistics in Medicine, 33(3), 517–535. https://doi.org/10.1002/sim.5941
[5]. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and Machine Learning. MIT Press. https://fairmlbook.org
[6]. Bhuiyan, M. I. H., & Mumu, T. B. (2022). Assessing the nexus between digital maturity and institutional accountability in Bangladesh's public health system: A 2022 cross-sectional analysis. *Journal of Medical and Health Studies*, *3*(4), 192–200. https://doi.org/10.32996/jmhs.2022.3.4.28
[7]. Chen, T., & Guestrin, C. (2016). XGBoost. Proceedings KDD'16, 785–794. https://doi.org/10.1145/2939672.2939785
[8]. Chouldechova, A. (2017). Fair prediction with disparate impact. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047
[9]. Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.
[10]. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608
[11]. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. NeurIPS 29.
[12]. Jayaprakash, S. M., et al. (2014). Early alert of academically at-risk students. Journal of Learning Analytics, 1(1), 6–47. https://doi.org/10.18608/jla.2014.11.2
[13]. Khosravi, H., et al. (2022). Explainable AI in education. Computers and Education: AI, 3, 100074. https://doi.org/10.1016/j.caeai.2022.100074
[14]. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). https://doi.org/10.4230/LIPIcs.ITCS.2017.43
[15]. Kuh, G. D., et al. (2006). What Matters to Student Success. NPEC.
[16]. Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57. https://doi.org/10.1145/3236386.3241340
[17]. Lonn, S., et al. (2012). Bridging the gap from knowledge to action. Proceedings LAK'12, 184–187. https://doi.org/10.1145/2330601.2330647
[18]. Lumley, T. (2010). Complex Surveys. Wiley. https://doi.org/10.1002/9780470580066
[19]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. NeurIPS 30.
[20]. Mitchell, M., et al. (2019). Model cards for model reporting. Proceedings FAccT'19, 220–229. https://doi.org/10.1145/3287560.3287596
[21]. NACADA. (2020). National Survey of Academic Advising. Kansas State University.
[22]. Namoun, A., & Alshanqiti, A. (2021). Predicting student performance. Applied Sciences, 11(1), 237. https://doi.org/10.3390/app11010237
[23]. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings ICML'05, 625–632. https://doi.org/10.1145/1102351.1102430
[24]. Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. International Statistical Review, 61(2), 317–337. https://doi.org/10.2307/1403631
[25]. Prenkaj, B., et al. (2021). A survey of ML for student dropout prediction. ACM Computing Surveys, 53(3), 1–34. https://doi.org/10.1145/3388792
[26]. Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical Linear Models (2nd ed.). Sage.
[27]. Reisman, D., et al. (2018). Algorithmic Impact Assessments. AI Now Institute.
[28]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?' Proceedings KDD'16, 1135–1144. https://doi.org/10.1145/2939672.2939778
[29]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
[30]. Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. Fordham Law Review, 87(5), 1085–1139.
[31]. Slack, D., et al. (2020). Fooling LIME and SHAP. Proceedings AIES'20, 180–186. https://doi.org/10.1145/3375627.3375830
[32]. Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). What are we weighting for? Journal of Human Resources, 50(2), 301–316. https://doi.org/10.3368/jhr.50.2.301
[33]. Steyerberg, E. W. (2019). Clinical Prediction Models (2nd ed.). Springer. https://doi.org/10.1007/978-3-030-16399-0
[34]. Tinto, V. (1987). Leaving College. University of Chicago Press.
[35]. Van Calster, B., et al. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230. https://doi.org/10.1186/s12916-019-1466-7
[36]. Wine, J., Cominole, M., & Wheeless, S. (2023). BPS:20/22 Data File Documentation (NCES 2023-102). NCES.
[37]. Wise, A. F. (2014). Designing pedagogical interventions. Proceedings LAK'14, 203–211. https://doi.org/10.1145/2567574.2567588