

TRUE-EWS: TRUSTWORTHY EARLY WARNING WITH EQUALIZED CONFORMAL GUARANTEES FOR EQUITABLE STUDENT SUPPORT IN U.S. EDUCATION

QianYi Fang

University of Chichester, Chichester PO19 6PE, United Kingdom.

Abstract: Early-warning systems are now routine in U.S. schools, colleges, and professional programs, yet a risk score alone does not tell a counselor how often at-risk students from each demographic group will be overlooked, or how often students who are on track will be flagged. This paper presents TRUE-EWS, a Transparent, Reliable, Uncertainty-aware, and Equitable early-warning framework. TRUE-EWS combines a gradient-boosted risk model, a strictly increasing group-wise Beta calibration map, an equalized conformal decision layer (ECC) that runs a Mondrian conformal predictor over group-by-outcome cells, and a TreeSHAP decomposition of between-group risk gaps. ECC maps each student to one of three actions (flag for support, no action, or referral to a counselor) and guarantees, in finite samples and without distributional assumptions, that the miss rate of at-risk students and the false-flag rate of students on track are bounded within every group. We also prove that strictly increasing recalibration leaves ECC decisions unchanged, so probability reliability and decision guarantees can be managed separately. Experiments with 30 stratified random splits on three U.S. cohorts that span the education pipeline, namely Tennessee STAR (early grades), High School and Beyond (college access), and the LSAC National Longitudinal Bar Passage Study (professional licensure), show that standard conformal prediction meets its 90% target overall while covering only 3.0% of at-risk White law students, and that label-conditional prediction flags 37.6% of Non-White students who go on to pass the bar. ECC holds every group-by-outcome coverage between 0.898 and 0.909, lowers the worst-group false-flag rate on LSAC from 0.371 to 0.098 relative to confidence-based deferral at the same counselor workload, and extends directly to intersectional groups.

Keywords: Trustworthy AI; Educational equity; Early warning systems; Conformal prediction; Algorithmic fairness; Probability calibration; Explainable AI; Human-in-the-loop decision making

1 INTRODUCTION

Predictive early-warning systems (EWS) have become part of the operating infrastructure of U.S. education. Districts flag students who may not graduate on time [1-2], universities score first-year students for dropout risk [3], and law schools identify students who may need support before the bar examination. The appeal is practical. Counseling, tutoring, and outreach capacity is scarce, and a well-placed flag moves help toward the students who need it before an adverse outcome occurs. The same systems operate on populations shaped by long-standing racial and socioeconomic gaps in achievement and access, and a model fitted to such data can carry those gaps into decisions about who receives attention [4-5].

Guidelines for trustworthy AI converge on transparency, justice and fairness, non-maleficence, responsibility, and privacy [6], and the AI-in-education community has called for frameworks that turn such principles into practice [7]. For an EWS, these principles translate into four operational questions that a district or university should be able to answer before deployment. Does a predicted risk of 30% mean the same thing for a Black student and a White student? What fraction of at-risk students in each group will the system fail to flag, and what fraction of students who are on track will it flag unnecessarily? When the evidence is weak, does the system say so, and who decides then? Which observable factors account for differences in predicted risk between groups? Research on fairness in education has concentrated on the second question for point predictions, through audits of error-rate gaps and through mitigation by reweighing [8-10], constrained optimization [11], or group-specific thresholds [12]. These interventions narrow error-rate gaps by giving up accuracy, and a classical impossibility result shows that a single thresholded score cannot be calibrated within groups and equalize error rates at the same time when base rates differ [13-14].

This paper takes a different route: it moves the fairness requirement from the score to the decision and gives the decision an explicit option to defer. TRUE-EWS keeps one risk model, calibrates it within groups so that the communicated probability is reliable, and routes every decision through an equalized conformal layer (ECC). ECC is a Mondrian conformal predictor whose calibration cells are indexed jointly by group membership and outcome label, which strengthens equalized coverage across groups to coverage within each group and outcome [15-17]. Its prediction set determines one of three actions: flag for support, no action, or referral to a counselor. Because coverage holds inside every cell, the miss rate of at-risk students and the false-flag rate of students on track are both bounded in every group, in finite samples and for any underlying model. Students for whom the evidence is ambiguous are handed to a person rather than forced into a label, in the spirit of learning to defer, and the deferral is controlled so that it does not widen

group disparities, as unconstrained selective classification can [18-19]. The familiar fairness–accuracy trade-off becomes an explicit fairness–workload trade-off that administrators can set to match counseling capacity.

The paper makes four contributions. (1) It proposes TRUE-EWS, an early-warning framework that addresses reliability, uncertainty, equity, and transparency with separate, composable components (Figure 1). (2) It establishes two properties of the ECC layer: finite-sample bounds on group-level miss and false-flag rates (Proposition 1), and invariance of ECC decisions to any strictly increasing group-wise recalibration (Proposition 2), which decouples communicated probabilities from decision guarantees and explains why piecewise-constant isotonic maps inflate referral workload. (3) It evaluates the framework on three public U.S. cohorts covering early grades, college access, and professional licensure, against point-decision fairness methods and four set-valued alternatives, including an intersectional extension. (4) It introduces an additive TreeSHAP decomposition that attributes the between-group gap in predicted risk to observable factors, which turns the model into an auditing instrument for equity planning.

Section 2 reviews related work. Section 3 formalizes the requirements, and Section 4 presents the framework and its guarantees. Sections 5 and 6 describe the experimental setup and results, Section 7 discusses implications for U.S. practice, and Section 8 concludes.

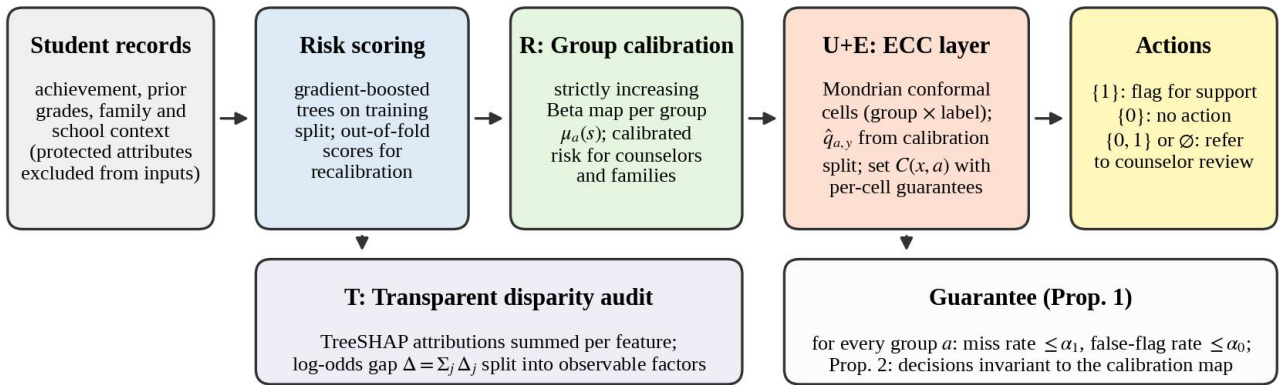


Figure 1 Overview of TRUE-EWS. A Risk Model Trained Without Protected Attributes Feeds a Strictly Increasing Group-Wise Calibration Map (Reliability); the Equalized Conformal Layer Computes Thresholds Within Group-by-Outcome Cells and Maps Each Prediction Set to an Action (Uncertainty Awareness and Equity); TreeSHAP Attributions Decompose Between-Group Risk Gaps (Transparency)

2 RELATED WORK

2.1 Early Warning and Predictive Analytics in U.S. Education

Early-warning research in the United States has a long empirical history. Bowers et al. compared 110 dropout flags from 36 studies with receiver operating characteristic analysis and found that most flags had high precision but limited accuracy, with longitudinal growth models and grade-based indicators performing best [1]. Lakkaraju et al. developed a machine learning framework with U.S. school districts to identify students at risk of not graduating on time and emphasized ranking students by risk so that schools can prioritize interventions [2]. In higher education, Yu et al. studied dropout prediction at a large public research university with residential and online students and found that adding protected attributes left overall performance essentially unchanged and only marginally improved fairness [3]. These studies establish that useful signal exists; the present work concerns the assurances that a deployed system should attach to that signal.

2.2 Algorithmic Fairness in Education

Baker and Hawn reviewed evidence of algorithmic bias across educational applications and demographic groups [4], and Kizilcec and Lee organized fairness notions and interventions for educational settings [5]. Gardner et al. introduced slicing analysis with the absolute between-ROC area to measure differential accuracy of student models [8], and Jiang and Pardos examined racial disparities in course-grade prediction together with mitigation strategies [9]. General-purpose interventions include reweighing training examples [10], exponentiated-gradient reductions that impose parity constraints [11], and group-specific thresholds that equalize odds after training [12]. Kleinberg et al. proved that calibration within groups and balance for both outcome classes cannot hold together outside degenerate cases [13], and Pleiss et al. showed that calibration is compatible with at most a single error constraint [14]. TRUE-EWS accepts this result and resolves it at the level of set-valued decisions rather than point decisions.

2.3 Calibration

A classifier is calibrated when its predicted probabilities match observed frequencies. Guo et al. measured miscalibration with the expected calibration error (ECE) and found temperature scaling, a one-parameter variant of Platt

scaling, to be effective [20]. Isotonic regression fits a nonparametric monotone map [21], and Beta calibration fits a parametric family derived from Beta-distributed class scores that contains the identity map and is monotone when its shape parameters are nonnegative [22]. For risk communication in schools, calibration within each group matters because a counselor who reads a risk estimate should not have to adjust it for a student's race or income.

2.4 Conformal Prediction, Equalized Coverage, and Deferral

Conformal prediction turns any predictive model into set-valued predictions with finite-sample coverage under exchangeability [15], and its split (inductive) form needs only a held-out calibration set. Vovk analyzed conditional versions of validity, including label-conditional and category-conditional (Mondrian) predictors [16]. Sadinle et al. derived least ambiguous set-valued classifiers with class-specific error control [23], and Romano et al. proposed equalized coverage, which guarantees equal coverage across protected groups [17]. On the decision side, Madras et al. showed that a model that learns when to defer to a downstream decision maker can improve both accuracy and fairness [18], whereas Jones et al. found that selective classification can magnify accuracy disparities between groups [19]. ECC combines group and label taxonomies inside one Mondrian predictor and uses the resulting sets as the decision layer of an EWS, with calibration and explanation as companion components. TreeSHAP provides exact Shapley-value attributions for tree ensembles such as LightGBM [24-25], which we aggregate into group-level disparity decompositions.

3 PROBLEM FORMULATION

Each student is described by features X , a group label A (for example race or ethnicity, possibly crossed with income), and a binary outcome $Y \in \{0,1\}$, where $Y = 1$ denotes the adverse outcome that support is meant to prevent. Protected attributes are not part of X . An early-warning policy $\delta(x,a)$ returns one of three actions: FLAG (proactive support), NONE (routine monitoring), or REFER (review by a counselor). We ask the policy to satisfy four requirements, one for each principle in the name of the framework.

R1 (Reliability). The communicated risk $\hat{p}(x,a)$ is calibrated within groups, i.e., $P(Y = 1 \mid \hat{p} = p, A = a)$ approx p for every group a .

R2 (Equity). For user-chosen levels α_1 and α_0 and every group a , the miss rate satisfies $P(\delta = \text{NONE} \mid Y = 1, A = a) \leq \alpha_1$ and the false-flag rate satisfies $P(\delta = \text{FLAG} \mid Y = 0, A = a) \leq \alpha_0$.

R3 (Uncertainty awareness). When the evidence supports neither automated action at the required error levels, the policy returns REFER.

R4 (Transparency). The gap in predicted log-odds between groups is decomposed exactly into contributions of observable features.

Requirement R2 is a finite-sample analogue of equalized odds that applies to automated actions [12]. With overlapping outcome distributions and small α_1 and α_0 , no binary classifier can meet both bounds, and the REFER option is what makes R2 attainable. The cost of R2 is the referral rate, which the framework reports by group.

4 THE TRUE-EWS FRAMEWORK

Figure 1 summarizes the framework and Algorithm 1 lists its steps. The available records are split into a training set D_{tr} and a calibration set D_{cal} .

4.1 Risk Scoring

A gradient-boosted decision tree (GBDT) ensemble f is trained on D_{tr} without protected attributes. We use LightGBM because it handles mixed and missing inputs and supports exact TreeSHAP attributions [25]. To fit calibration maps without reusing in-sample scores, out-of-fold scores s_i are computed on D_{tr} by five-fold cross-fitting, while the deployed model f is trained on all of D_{tr} .

4.2 Group-Wise Beta Calibration

For each group a we fit the Beta calibration map [22]

$$\mu_a(s) = \sigma(c_a + a_a \ln s - b_a \ln(1 - s)) \quad (1)$$

where $\sigma(t) = 1/(1 + e^{-t})$, by logistic regression of Y on $(\ln s, -\ln(1 - s))$ using the out-of-fold scores of students in group a . When an estimated slope is negative, the corresponding term is dropped and the map is refit, following [22]. Because $d\mu_a/ds = \mu_a(1 - \mu_a)(a_a/s + b_a/(1 - s))$, the map is strictly increasing whenever $a_a, b_a \geq 0$ and $a_a + b_a > 0$. The communicated risk is $\hat{p}(x,a) = \mu_a(f(x))$.

4.3 Equalized Conformal Decision Layer (ECC)

Let $\hat{p}_1(x,a) = \hat{p}(x,a)$ and $\hat{p}_0(x,a) = 1 - \hat{p}(x,a)$. ECC uses the least-ambiguous nonconformity score [23]:

$$V(x,a,y) = 1 - \hat{p}_y(x,a) \quad (2)$$

and partitions D_{cal} into cells $I_{a,y} = \{i : A_i = a, Y_i = y\}$ of size $n_{a,y}$. The threshold of each cell is the $k_{a,y}$ -th smallest calibration score,

$$\hat{q}_{a,y} = V_{(k_{a,y})}, k_{a,y} = \lceil (n_{a,y} + 1)(1 - \alpha_y) \rceil \quad (3)$$

with $\hat{q}_{a,y} = +\infty$ when $k_{a,y} > n_{a,y}$. For a new student with features x in group a , the prediction set and the action are

$$C(x,a) = \{y \in \{0,1\} : V(x,a,y) \leq \hat{q}_{a,y}\} \quad (4)$$

$$\delta(x,a) = \begin{cases} \text{FLAG}, & \text{if } C(x,a) = \{1\}, \\ \text{NONE}, & \text{if } C(x,a) = \{0\}, \\ \text{REFER}, & \text{otherwise.} \end{cases} \quad (5)$$

REFER covers both the ambiguous set $\{0,1\}$ and the empty set, which marks a student who is atypical for both outcomes in that group. Separate levels $\alpha_1 \leq \alpha_0$ encode the common judgment that missing an at-risk student costs more than an unnecessary check-in.

Proposition 1 (Group–outcome validity). If, for each cell (a,y) , the calibration examples in $I_{a,y}$ and a test example from the same cell are exchangeable, then $P(y \in C(X,A) \mid A = a, Y = y) \geq 1 - \alpha_y$. Consequently, for every group a , $P(\delta = \text{NONE} \mid Y = 1, A = a) \leq \alpha_1$ and $P(\delta = \text{FLAG} \mid Y = 0, A = a) \leq \alpha_0$. If the scores are almost surely distinct, the coverage is also below $1 - \alpha_y + 1/(n_{a,y} + 1)$.

Proof. Conditional on the cell, the test score $V(X,a,y)$ and the $n_{a,y}$ calibration scores are exchangeable, so the rank of the test score among the $n_{a,y} + 1$ values is uniformly distributed when ties are broken at random. The event $V(X,a,y) \leq \hat{q}_{a,y}$ therefore has probability at least $k_{a,y}/(n_{a,y} + 1) \geq 1 - \alpha_y$, and exactly $k_{a,y}/(n_{a,y} + 1) < 1 - \alpha_y + 1/(n_{a,y} + 1)$ without ties [15-16]. An at-risk student receives NONE only if $1 \notin C$, and a student on track receives FLAG only if $0 \notin C$, which gives the two bounds. ■

Proposition 2 (Invariance to strictly increasing recalibration). Let g_a be strictly increasing on $(0,1)$ for every group a . The ECC sets computed from raw scores $s = f(x)$ and from recalibrated scores $g_a(s)$, with the same calibration data and levels, coincide for every x and a .

Proof. For $y = 1$ the score $1 - g_a(s)$ is a strictly decreasing function of s , exactly as $1 - s$ is, so within cell $(a,1)$ the ordering of the calibration and test scores is identical under both representations, and the test score lies at or below the $k_{a,1}$ -th order statistic in both cases or in neither. The same argument with increasing functions applies to $y = 0$, whose score is $g_a(s)$. ■

Proposition 2 has two practical consequences. First, calibration maps can be refreshed, for instance every term, without changing any decision, so the reliability of the communicated risk and the validity of the actions are maintained independently. Second, maps that are monotone but not strictly increasing, such as isotonic regression [21], produce piecewise-constant scores; ties at a threshold are admitted together, which enlarges prediction sets and raises the referral workload without improving validity. Section 6.6 quantifies this effect.

4.4 Transparent Disparity Audit

Let $f(x) = \phi_0 + \sum_j \phi_j(x)$ be the TreeSHAP decomposition of the GBDT log-odds [24], where the one-hot columns of an original feature are summed into a single ϕ_j . Because the base value ϕ_0 cancels, the between-group gap in predicted log-odds decomposes exactly as

$$\Delta = E[f(X) \mid A = 1] - E[f(X) \mid A = 0] = \sum_j \Delta_j \quad (6)$$

with $\Delta_j = E[\phi_j(X) \mid A = 1] - E[\phi_j(X) \mid A = 0]$. A large contribution from prior achievement indicates that the higher predicted risk of a group is carried by an achievement gap that support programs can target, whereas a large contribution from a contextual variable points to structural access conditions. Both are reported together with group-level miss, false-flag, and referral rates.

Algorithm 1: TRUE-EWS

Input: training set D_{tr} , calibration set D_{cal} , levels α_1, α_0

1: Train the GBDT f on D_{tr} using features X only.

2: Compute five-fold out-of-fold scores on D_{tr} ; fit μ_a for every group a by (1).

3: For each cell (a,y) of D_{cal} , compute V_i by (2) and $\hat{q}_{a,y}$ by (3).

4: For a new student (x,a) : report $\hat{p} = \mu_a(f(x))$, form $C(x,a)$ by (4), and return $\delta(x,a)$ by (5).

5: Each term, audit Δ_j by (6) together with group-level miss, false-flag, and referral rates.

5 EXPERIMENTAL SETUP

5.1 Datasets

We use three public U.S. cohorts that represent distinct stages of the education pipeline (Table 1).

Tennessee STAR. Project STAR randomly assigned about 11,600 Tennessee students and their teachers to small classes, regular classes, or regular classes with an aide from kindergarten through grade 3, beginning in 1985–86 [26]. We use the public release in [27] and place the early-warning point at the end of grade 1. Inputs are kindergarten and grade-1 reading and mathematics scaled scores (kindergarten scores are missing for students who joined in grade 1 and are encoded with missing indicators), birth date, class type in kindergarten and grade 1, school urbanicity, and the grade-1 teacher's degree, career-ladder level, experience, and race. A student is at risk if the sum of standardized grade-3 reading and mathematics scores falls in the bottom quartile. The groups are Black students versus all others (A) and

free-lunch eligibility in grade 1 (S). The analytic sample contains the 3,762 students with grade-1 and grade-3 scores and recorded lunch status.

High School and Beyond (HS&B). This extract of the 1980 senior cohort with a 1986 follow-up was assembled by Rouse and distributed in [27–28]. Inputs are a base-year composite achievement score, whether the father and the mother graduated from college, home ownership, urban location, county unemployment rate, state hourly manufacturing wage, distance to the nearest four-year college, average state four-year tuition, and region. A student is at risk if no schooling beyond high school (12 years) is reported at follow-up. The groups are African American or Hispanic students (URM) versus others (A) and family income below USD 25,000 (S); $n = 4,739$.

LSAC Bar Passage. The LSAC National Longitudinal Bar Passage Study followed students who entered ABA-approved law schools in fall 1991 through their bar examinations [29]. We use the cleaned release of 20,798 students described in [30]. Inputs are those available after the first year: first-year GPA (z-score), first-year class-rank decile, LSAT score, undergraduate GPA, family-income category, full-time status, and law-school tier. A student is at risk if the bar examination was not passed. The groups are Non-White versus White students (A) and women versus men (S).

The at-risk prevalence differs across groups on every dataset (Table 1), from 0.079 versus 0.277 on LSAC to 0.370 versus 0.416 on HS&B, which creates very different conditions for the decision layer.

Table 1 Datasets and Early-Warning Tasks

	STAR	HS&B	LSAC
Stage	Grades K–3	College access	Licensure
Cohort	1985–1989	1980–1986	1991 entrants
Students n	3,762	4,739	20,798
Inputs (encoded)	12 (31)	10 (15)	7 (13)
At-risk outcome	Bottom-quartile grade-3 score	No postsecondary schooling	Bar exam not passed
$P(Y = 1)$	0.250	0.387	0.110
Group $A = 1$ (share)	Black (0.285)	URM (0.356)	Non-White (0.159)
$P(Y = 1 \mid A = 0 / 1)$	0.184 / 0.415	0.370 / 0.416	0.079 / 0.277
Group $S = 1$ (share)	Free lunch (0.427)	Low income (0.712)	Female (0.439)

5.2 Protocol and Baselines

Each experiment is repeated over 30 random splits stratified by (A, Y) into 60% training, 20% calibration, and 20% test data, and means over the splits are reported. Protected attributes never enter the models; they define calibration cells and are used by the fairness baselines and in evaluation. The GBDT uses 300 trees, learning rate 0.03, 15 leaves, at least 30 samples per leaf, row and column subsampling of 0.8, and an L2 penalty of 1. Logistic regression (LR) uses standardized inputs.

Point-decision baselines are LR, GBDT, GBDT with reweighing [10] (GBDT-RW, weights $w(a, y) = P(a)P(y)/P(a, y)$), the exponentiated-gradient reduction [11] with an equalized-odds constraint (bound 0.02), cost-sensitive error weights $(\pi, 1 - \pi)$ for training prevalence π , and an LR base learner (ExpGrad-EO), and randomized group-specific thresholds [12] fitted on the calibration split to maximize balanced accuracy under equalized odds (ThreshOpt-EO). Score-based models flag a student when the score is at least π .

Set-valued baselines use the same calibrated scores as TRUE-EWS: marginal split conformal prediction with the score in (2); group-conditional prediction, i.e., equalized coverage [17]; label-conditional prediction [16, 23]; and confidence-based deferral (SR-defer) in the sense of [19], which refers the students with the smallest $|\text{logit } \hat{p} - \text{logit } \pi|$ using one global cut-off chosen so that its overall referral rate equals that of ECC, and flags the remaining students with $\hat{p} \geq \pi$. Unless stated otherwise, $\alpha_1 = \alpha_0 = 0.1$.

5.3 Metrics

We report the area under the ROC curve (AUC), balanced accuracy (BAcc), the equal-opportunity gap $\Delta\text{TPR} = |\text{TPR}_1 - \text{TPR}_0|$, and the equalized-odds gap $\Delta\text{EO} = \max(\Delta\text{TPR}, \Delta\text{FPR})$. Calibration is measured by ECE with ten equal-width bins, overall and within each group. For set-valued decisions we report the cell coverage $\text{Cov}(a, y) = P(y \in C \mid A = a, Y = y)$, the referral rate, the worst-group miss rate $\max_a P(\delta = \text{NONE} \mid Y = 1, A = a)$, and the worst-group false-flag rate $\max_a P(\delta = \text{FLAG} \mid Y = 0, A = a)$, computed from rates averaged over the splits. For point classifiers the miss and false-flag rates are $1 - \text{TPR}$ and FPR .

6 RESULTS

6.1 Discrimination and Point-Decision Fairness

Table 2 Discrimination and Point-Decision Fairness (Mean over 30 Splits)

Method	AUC	BAcc	ΔTPR	ΔEO	Miss
Tennessee STAR					
LR	0.862	0.781	0.146	0.253	0.272

Method	AUC	BAcc	Δ TPR	Δ EO	Miss
GBDT	0.863	0.780	0.137	0.244	0.284
GBDT-RW [10]	0.857	0.777	0.063	0.155	0.251
ExpGrad-EO [11]	0.762	0.741	0.070	0.084	0.275
ThreshOpt-EO [12]	—	0.740	0.071	0.084	0.277
TRUE-EWS score	0.864	0.782	0.150	0.318	0.248
High School and Beyond					
LR	0.753	0.684	0.210	0.278	0.387
GBDT	0.740	0.677	0.171	0.205	0.398
GBDT-RW [10]	0.740	0.675	0.157	0.185	0.395
ExpGrad-EO [11]	0.665	0.652	0.039	0.067	0.342
ThreshOpt-EO [12]	—	0.672	0.047	0.068	0.339
TRUE-EWS score	0.741	0.677	0.123	0.154	0.353
LSAC Bar Passage					
LR	0.847	0.767	0.242	0.340	0.320
GBDT	0.842	0.761	0.236	0.336	0.327
GBDT-RW [10]	0.835	0.755	0.172	0.297	0.272
ExpGrad-EO [11]	0.707	0.669	0.046	0.051	0.330
ThreshOpt-EO [12]	—	0.723	0.042	0.049	0.257
TRUE-EWS score	0.843	0.763	0.265	0.424	0.318

Note: Miss is the worst-group share of at-risk students who are not flagged. Standard deviations over splits are at most 0.03 for AUC and BAcc and at most 0.07 for Δ TPR and Δ EO.

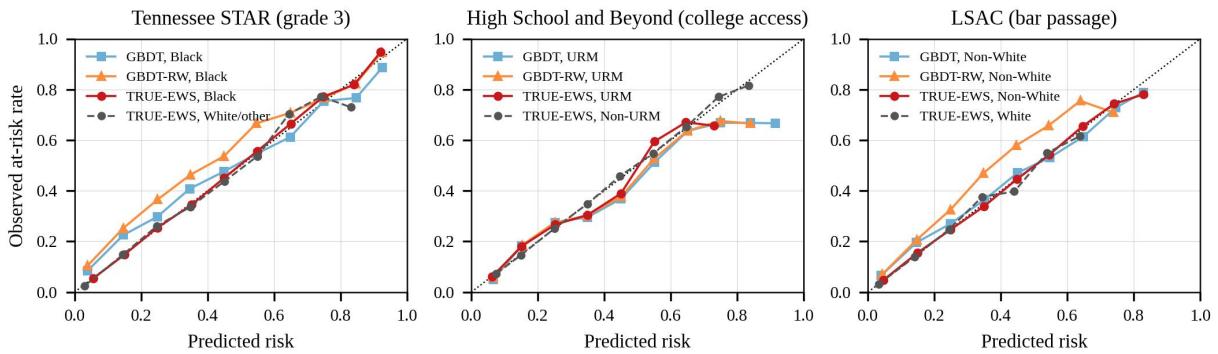


Figure 2 Reliability Diagrams on the Pooled Test Splits. Points Above the Diagonal Indicate Underestimated Risk. Reweighting (GBDT-RW) Understates the Risk of Black (STAR) and Non-White (LSAC) Students, Whereas Group-Wise Beta Calibration (TRUE-EWS) Aligns Both Groups With the Diagonal

Table 2 reports discrimination and error-rate gaps for point decisions. LR and GBDT reach similar AUC on every dataset (0.863 and 0.862 on STAR, 0.740 and 0.753 on HS&B, 0.842 and 0.847 on LSAC), and both leave large gaps between groups: the unconstrained GBDT has Δ EO of 0.244, 0.205, and 0.336. Base-rate differences drive these gaps. On LSAC, flagging every student whose score exceeds the overall prevalence flags 53.8% of the Non-White students who later pass the bar, compared with 20.2% of White students. Reweighting narrows Δ TPR on STAR (0.137 to 0.063) and LSAC (0.236 to 0.172) at a cost of at most 0.007 in AUC. ExpGrad-EO and ThreshOpt-EO bring Δ EO down to between 0.049 and 0.084, but they lower BAcc by up to 0.092 and still leave 25.7% to 34.2% of the at-risk students in the worse-served group without a flag.

The scoring layer of TRUE-EWS (GBDT with group-wise Beta calibration) keeps the AUC of the GBDT (0.864, 0.741, 0.843). Thresholding it widens Δ EO on STAR (0.318) and LSAC (0.424), the two datasets with the largest base-rate differences, and narrows it on HS&B (0.154). This is the behavior that predict for scores calibrated within groups [13–14], and it is the reason TRUE-EWS does not threshold its score but passes it to ECC.

6.2 Reliability of Communicated Risk

All GBDT-based score layers are close to calibrated overall, with ECE between 0.008 and 0.046, yet calibration is uneven across groups (Table 3). For the unaware GBDT the worst-group ECE is 0.081 on STAR, 0.076 on HS&B, and 0.041 on LSAC, and the higher-prevalence group is the worse-calibrated one in every case. Reweighting aggravates the problem on STAR (0.098) and LSAC (0.070): its weights reduce the influence of at-risk students in the group with the higher base rate, and Figure 2 shows the resulting underestimation of risk for Black and Non-White students. Group-wise Beta calibration reduces the worst-group ECE to 0.065, 0.060, and 0.035, which is below a single global Beta map on every dataset and within 0.003 of group-wise isotonic regression, while the map remains strictly increasing. After recalibration the curves of both groups follow the diagonal in Figure 2.

Table 3 Calibration of Communicated Risk: Overall and Worst-Group ECE

	STAR		HS&B		LSAC	
Score layer	All	Worst	All	Worst	All	Worst

	STAR		HS&B		LSAC	
GBDT (raw)	0.040	0.081	0.046	0.076	0.010	0.041
GBDT-RW (raw)	0.039	0.098	0.045	0.071	0.009	0.070
+ Global Beta	0.029	0.070	0.032	0.062	0.008	0.039
+ Group isotonic	0.031	0.064	0.034	0.057	0.009	0.035
+ Group Beta (ours)	0.029	0.065	0.032	0.060	0.008	0.035

Note: Worst is the larger of the two within-group ECE values, averaged over splits.

6.3 Decision-Layer Guarantees

Table 4 and Figure 3 compare the decision layers at $\alpha = 0.1$. Every method reaches about 90% overall coverage, but the methods distribute their errors very differently. Marginal conformal prediction covers only 60.8% of at-risk White/other students on STAR, 76.8% of at-risk non-URM students on HS&B, and 3.0% of at-risk White students on LSAC. Under class imbalance the cheapest way to reach 90% marginal coverage is to output $\{0\}$ for most students, and on LSAC the method refers only 0.7% of them. Group-conditional prediction equalizes coverage across groups by design [17], but conditioning on the group alone does not protect at-risk students; on LSAC it covers 0.1% of at-risk White students.

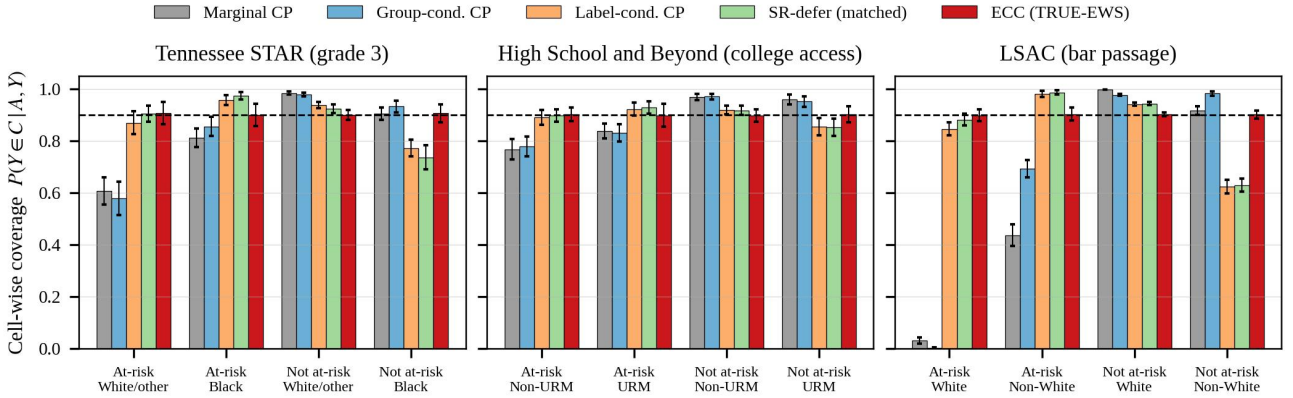


Figure 3 Cell-Wise Coverage at $\alpha = 0.1$ (Mean $\pm$ s.d. over 30 Splits); the Dashed Line Marks the 0.9 Target. Only ECC Reaches the Target in All Four Group-by-Outcome Cells on Every Dataset

Table 4 Cell-Wise Coverage and Referral Rate at $\alpha = 0.1$

	At-risk ($Y = 1$)		On-track ($Y = 0$)		
Method	A = 0	A = 1	A = 0	A = 1	Referral
Tennessee STAR					
Marginal CP	0.608	0.813	0.985	0.905	0.201
Group-cond. CP [17]	0.579	0.857	0.980	0.933	0.215
Label-cond. CP [16]	0.871	0.958	0.939	0.773	0.292
SR-defer [19]	0.906	0.975	0.924	0.737	0.316
ECC (ours)	0.909	0.901	0.901	0.907	0.316
High School and Beyond					
Marginal CP	0.768	0.839	0.970	0.960	0.519
Group-cond. CP [17]	0.779	0.831	0.972	0.952	0.520
Label-cond. CP [16]	0.891	0.923	0.920	0.856	0.545
SR-defer [19]	0.899	0.929	0.918	0.853	0.554
ECC (ours)	0.903	0.899	0.898	0.903	0.554
LSAC Bar Passage					
Marginal CP	0.030	0.436	0.999	0.917	0.007
Group-cond. CP [17]	0.001	0.693	0.978	0.983	0.082
Label-cond. CP [16]	0.847	0.982	0.942	0.624	0.333
SR-defer [19]	0.883	0.987	0.944	0.629	0.385
ECC (ours)	0.899	0.904	0.903	0.902	0.385

Note: Entries are $\text{Cov}(a, y)$ averaged over 30 splits; standard deviations are at most 0.07.

Label-conditional prediction restores the coverage of at-risk students (0.847 to 0.982) and moves the errors onto on-track students in the higher-prevalence group: it covers only 77.3% of on-track Black students on STAR, 85.6% of on-track URM students on HS&B, and 62.4% of on-track Non-White students on LSAC, and the rest are flagged. SR-defer shows the same pattern at the referral budget of ECC (on-track coverage of 0.737, 0.853, and 0.629 in that group), in line with the observation that selective classification can magnify disparities [19]. ECC is the only method that keeps all twelve cells between 0.898 and 0.909.

The referral rates of ECC are 0.316 on STAR, 0.554 on HS&B, and 0.385 on LSAC, and they are higher for the group with the higher base rate (0.433 versus 0.269, 0.593 versus 0.533, and 0.447 versus 0.373). Counselor attention is

therefore directed to where the evidence of the model is weakest, and the allocation is explicit and auditable instead of being absorbed into unequal error rates.

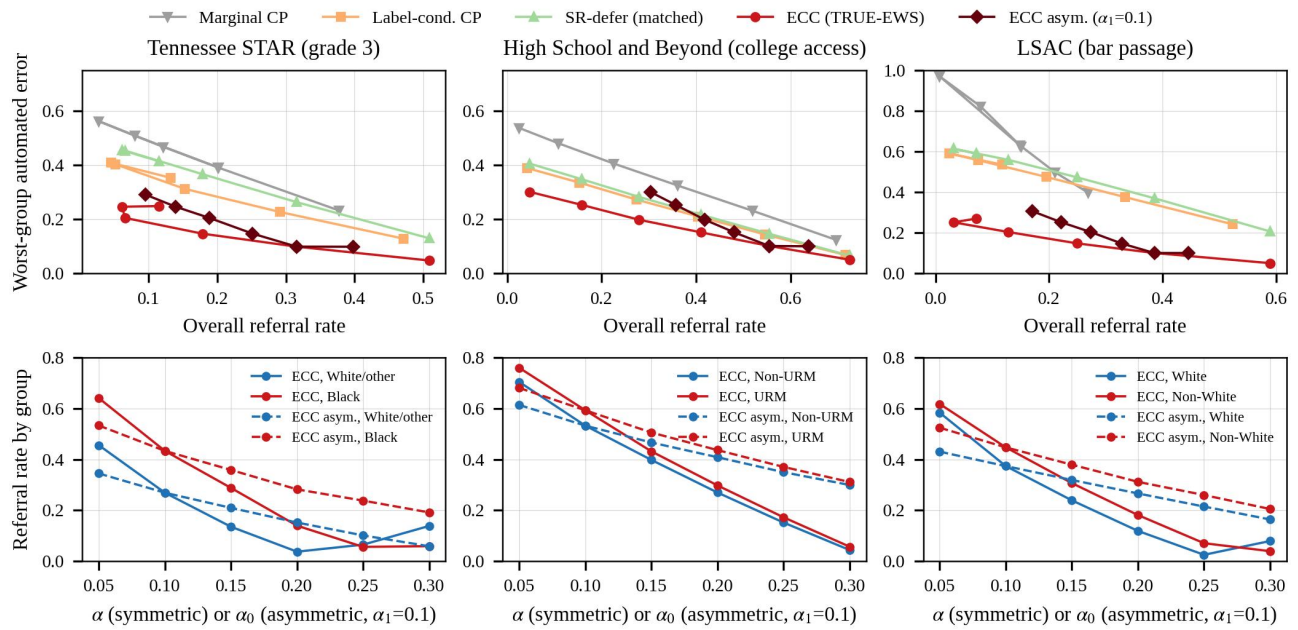


Figure 4 Fairness–Workload Trade-off. Top: Worst-Group Automated Error (the Largest Group-Level Miss or False-Flag Rate) Versus the Overall Referral Rate as α Varies from 0.05 to 0.30; SR-defer Is Evaluated at the Referral Budget of ECC. Bottom: Referral Rate by Group for Symmetric ECC (Solid) and Asymmetric ECC with $\alpha_1 = 0.1$ (Dashed)

6.4 Fairness–Workload Trade-off

Table 5 Worst-Group Automated Error and Counselor Workload at $\alpha = 0.1$

Policy	Worst miss	Worst false flag	Referral
Tennessee STAR			
GBDT	0.284	0.410	0
ExpGrad-EO	0.275	0.296	0
ThreshOpt-EO	0.277	0.281	0
Label-cond. CP	0.129	0.227	0.292
SR-defer	0.094	0.263	0.316
ECC (ours)	0.099	0.099	0.316
High School and Beyond			
GBDT	0.398	0.441	0
ExpGrad-EO	0.342	0.394	0
ThreshOpt-EO	0.339	0.324	0
Label-cond. CP	0.109	0.144	0.545
SR-defer	0.101	0.147	0.554
ECC (ours)	0.101	0.102	0.554
LSAC Bar Passage			
GBDT	0.327	0.538	0
ExpGrad-EO	0.330	0.376	0
ThreshOpt-EO	0.257	0.310	0
Label-cond. CP	0.153	0.376	0.333
SR-defer	0.117	0.371	0.385
ECC (ours)	0.101	0.098	0.385

Table 5 places point and set-valued policies on a common scale: the worst-group miss rate, the worst-group false-flag rate, and the referral rate. Point classifiers refer nobody and pay with large automated errors. The fairness-constrained ones still miss 25.7% to 34.2% of at-risk students in the worse-served group, and every point classifier flags at least 28.1% of on-track students in some group. ECC keeps both worst-group rates between 0.098 and 0.102 on all datasets while referring 31.6% to 55.4% of students. At the same referral budget, SR-defer reaches worst-group false-flag rates of 0.263, 0.147, and 0.371.

Figure 4 extends the comparison over levels from 0.05 to 0.30. On all three datasets the curve of ECC lies below those of the other decision layers, and its worst-group automated error tracks the chosen level. Lowering the workload requires accepting larger errors, and the asymmetric variant lets administrators decide where. With α_1 fixed at 0.1 and α_0

raised to 0.3, the referral rate falls from 0.316 to 0.096 on STAR, from 0.554 to 0.303 on HS&B, and from 0.385 to 0.170 on LSAC, while the worst-group miss rate stays at 0.099, 0.101, and 0.101. HS&B needs the most referrals because its inputs carry the weakest signal (AUC 0.740); ECC reports this as workload instead of converting it into unequal errors.

6.5 Intersectional Groups

Table 6 Intersectional Groups: Minimum Cell Coverage over Four Groups

Method	Min Cov at-risk	Min Cov on-track	Referral
Tennessee STAR (race $\times$ free lunch)			
Marginal CP	0.490	0.865	0.203
Group-cond. CP	0.356	0.905	0.216
Label-cond. CP	0.761	0.676	0.288
ECC (ours)	0.913	0.898	0.373
High School and Beyond (URM $\times$ income)			
Marginal CP	0.503	0.944	0.507
Group-cond. CP	0.599	0.942	0.516
Label-cond. CP	0.669	0.837	0.534
ECC (ours)	0.900	0.887	0.582
LSAC Bar Passage (race $\times$ sex)			
Marginal CP	0.022	0.915	0.007
Group-cond. CP	0.000	0.975	0.082
Label-cond. CP	0.845	0.583	0.331
ECC (ours)	0.898	0.900	0.392

Equity questions in U.S. education often concern the intersection of race with income or gender. Crossing A with S gives four groups per dataset and eight calibration cells. Table 6 shows that the failure modes of the alternatives become sharper at this resolution. Marginal prediction covers 2.2% of at-risk White men on LSAC and 49.0% of at-risk White/other students without free lunch on STAR, and label-conditional prediction covers only 58.3% of on-track Non-White women on LSAC. ECC with intersectional cells keeps every one of the 24 cells between 0.887 and 0.939. The overall referral rate rises moderately relative to two groups (0.373 versus 0.316 on STAR, 0.582 versus 0.554 on HS&B, and 0.392 versus 0.385 on LSAC).

6.6 Ablation

Table 7 isolates the components of TRUE-EWS. Removing calibration or replacing the group-wise map with a global one leaves every decision unchanged, exactly as Proposition 2 predicts, and only worsens the worst-group ECE (from 0.065 to 0.081 on STAR without calibration). Replacing the Beta map with group-wise isotonic regression gives a similar ECE but raises the referral rate by 6.3 to 9.1 percentage points, because tied scores at the thresholds are admitted together and the sets become conservative, with worst-group error rates of 0.075 to 0.082, below the target. Training the base model with reweighing changes the referral rate by at most 0.006 and leaves the guarantees intact, so the decision layer does not depend on a fairness-constrained learner. The guarantees come from the group-by-outcome conditioning itself: dropping the group index gives label-conditional prediction and dropping the outcome index gives group-conditional prediction, whose failures appear in Table 4.

Table 7 Ablation of TRUE-EWS Components at $\alpha = 0.1$

Variant	Worst ECE	Referral	Worst miss	Worst false flag
Tennessee STAR				
TRUE-EWS (full)	0.065	0.316	0.099	0.099
Reweighted base	0.061	0.322	0.097	0.100
Global Beta map	0.070	0.316	0.099	0.099
No calibration	0.081	0.316	0.099	0.099
Group isotonic map	0.064	0.382	0.075	0.080
High School and Beyond				
TRUE-EWS (full)	0.060	0.554	0.101	0.102
Reweighted base	0.062	0.554	0.102	0.103
Global Beta map	0.062	0.554	0.101	0.102
No calibration	0.076	0.554	0.101	0.102
Group isotonic map	0.057	0.645	0.075	0.082
LSAC Bar Passage				
TRUE-EWS (full)	0.035	0.385	0.101	0.098
Reweighted base	0.031	0.390	0.102	0.102
Global Beta map	0.039	0.385	0.101	0.098
No calibration	0.041	0.385	0.101	0.098
Group isotonic map	0.035	0.448	0.081	0.081

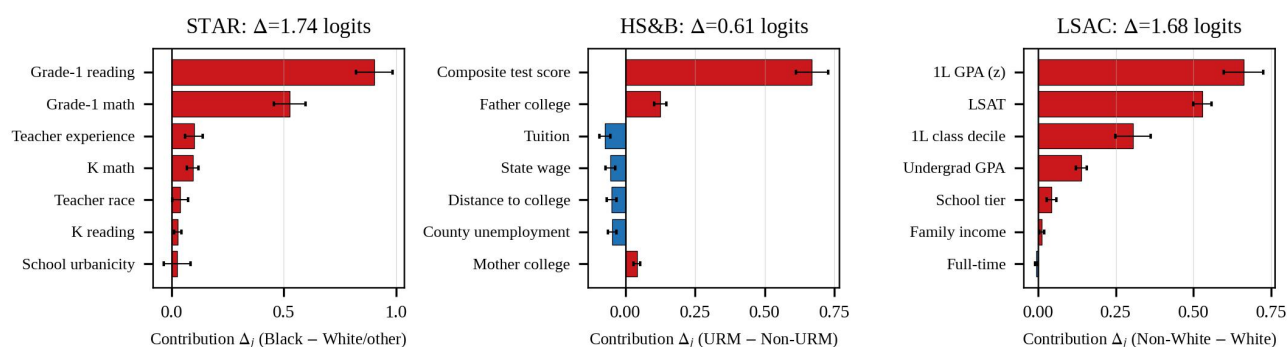


Figure 5 TreeSHAP Decomposition of the Between-Group Gap in Predicted Log-Odds ($\Delta = \sum_j \Delta_j$; Mean $\pm$ s.d. over Test Splits). Positive Contributions Raise the Predicted Risk of the Group Listed First in the Axis Label

6.7 Transparent Disparity Audit

Figure 5 decomposes the between-group gap in predicted log-odds with (6). On STAR the gap between Black and White/other students is 1.74 logits, and grade-1 reading and mathematics account for 1.43 of it (82%), while class type contributes at most 0.01 and school urbanicity 0.03. The higher predicted risk of Black students is therefore carried by measured achievement gaps at the end of grade 1, which directs support toward early literacy and numeracy. On LSAC, first-year GPA, LSAT score, first-year class rank, and undergraduate GPA account for 97% of a 1.68-logit gap, and family income contributes 0.01. On HS&B the composite test score contributes 0.67 logits, more than the total gap of 0.61, parental college education adds 0.17, and the contextual access variables (tuition, state wage, distance to a four-year college, and county unemployment) offset 0.23 logits. In this cohort the predicted access gap for URM students is driven by academic preparation and parental education, while their local access conditions lower their predicted risk.

7 DISCUSSION

The results carry several implications for U.S. schools, colleges, and professional programs that deploy early-warning analytics.

Capacity planning. The referral curves in Figure 4 translate directly into staffing. A district that can review about one fifth of a grade-1 cohort can set $\alpha_1 = 0.1$ and $\alpha_0 = 0.2$ and obtain a referral rate of 0.189 on STAR while the worst-group miss rate stays at 0.099. Because the guarantee is stated per group, it can be checked every term by counting outcomes.

Where fairness lives. Fairness constraints on point classifiers equalize error rates by giving up accuracy, and group calibration of a single thresholded score cannot equalize error rates when base rates differ. TRUE-EWS places calibration in the communicated risk and equity in the actions, and Proposition 2 ensures that the two do not interfere. Group membership is used only to set the thresholds that allocate support, and the resulting referral rates are reported by group.

Signal strength across the pipeline. The guarantees hold on all three cohorts, and the workload they require reflects the information available at each decision point. Early achievement scores on STAR and first-year law-school grades on LSAC carry strong signal, whereas the HS&B inputs are coarse. Richer inputs such as attendance, course grades, and financial-aid milestones would lower the referral rate without any change to the validity argument, which holds for any score model.

Auditing equity. The disparity decomposition turns the model into a measurement tool. On STAR and LSAC, predicted gaps are carried by academic achievement measured earlier in the pipeline; on HS&B, local access conditions work in favor of URM students while preparation works against them. Such findings indicate where upstream investment would reduce later risk. Future work will extend ECC to cohort-to-cohort distribution shift, to multi-level outcomes such as course grades, and to settings in which counselor decisions on referred students feed back into model updates.

8 CONCLUSION

TRUE-EWS shows that an early-warning system can give schools and universities concrete, checkable assurances. Calibration within groups makes a communicated risk mean the same thing for every student; the equalized conformal layer bounds, for every group, how often at-risk students are missed and how often students on track are flagged; uncertain cases go to counselors in a quantity that administrators control; and a TreeSHAP decomposition explains what drives risk gaps. Across three U.S. cohorts from early grades to professional licensure, ECC held every group-by-outcome coverage between 0.898 and 0.909, whereas each marginal, group-conditional, label-conditional, and confidence-based alternative left at least one cell at or below 0.629. These properties make the framework a practical basis for trustworthy AI that widens access to support and improves student outcomes.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] Bowers A J, Sprott R, Taff S A. Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, 2013, 96(2): 77-100.
- [2] Lakkaraju H, Aguiar E, Shan C, et al. A machine learning framework to identify students at risk of adverse academic outcomes. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015: 1909-1918.
- [3] Yu R, Lee H, Kizilcec R F. Should college dropout prediction models include protected attributes? *Proceedings of the 8th ACM Conference on Learning @ Scale*, 2021: 91-100.
- [4] Baker R S, Hawn A. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 2022, 32(4): 1052-1092.
- [5] Kizilcec R F, Lee H. Algorithmic fairness in education. In: Holmes W, Porayska-Pomsta K, eds. *The Ethics of Artificial Intelligence in Education*. Routledge, 2022: 174-202.
- [6] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 2019, 1(9): 389-399.
- [7] Holmes W, Porayska-Pomsta K, Holstein K, et al. Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 2022, 32(3): 504-526.
- [8] Gardner J, Brooks C, Baker R. Evaluating the fairness of predictive student models through slicing analysis. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 2019: 225-234.
- [9] Jiang W, Pardos Z A. Towards equity and algorithmic fairness in student grade prediction. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2021: 608-617.
- [10] Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 2012, 33(1): 1-33.
- [11] Agarwal A, Beygelzimer A, Dudík M, et al. A reductions approach to fair classification. *Proceedings of the 35th International Conference on Machine Learning*, PMLR, 2018, 80: 60-69.
- [12] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 2016, 29: 3315-3323.
- [13] Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017: 43:1-43:23.
- [14] Pleiss G, Raghavan M, Wu F, et al. On fairness and calibration. *Advances in Neural Information Processing Systems*, 2017, 30: 5680-5689.
- [15] Vovk V, Gammelman A, Shafer G. *Algorithmic Learning in a Random World*. New York: Springer, 2005.
- [16] Vovk V. Conditional validity of inductive conformal predictors. *Proceedings of the Asian Conference on Machine Learning*, PMLR, 2012, 25: 475-490.
- [17] Romano Y, Barber R F, Sabatti C, et al. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2020, 2(2).
- [18] Madras D, Pitassi T, Zemel R. Predict responsibly: Improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 2018, 31.
- [19] Jones E, Sagawa S, Koh P W, et al. Selective classification can magnify disparities across groups. *Proceedings of the International Conference on Learning Representations*, 2021.
- [20] Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, PMLR, 2017, 70: 1321-1330.
- [21] Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002: 694-699.
- [22] Kull M, Silva Filho T, Flach P. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, PMLR, 2017, 54: 623-631.
- [23] Sadinle M, Lei J, Wasserman L. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 2019, 114(525): 223-234.
- [24] Lundberg S M, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2020, 2(1): 56-67.
- [25] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 2017, 30: 3146-3154.
- [26] Krueger A B. Experimental estimates of education production functions. *Quarterly Journal of Economics*, 1999, 114(2): 497-532.
- [27] Kleiber C, Zeileis A. *Applied Econometrics with R*. New York: Springer, 2008.
- [28] Rouse C E. Democratization or diversion? The effect of community colleges on educational attainment. *Journal of Business & Economic Statistics*, 1995, 13(2): 217-224.
- [29] Wightman L F. *LSAC National Longitudinal Bar Passage Study*. Newtown: Law School Admission Council, LSAC Research Report Series, 1998.
- [30] Le Quy T, Roy A, Iosifidis V, et al. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, 2022, 12(3): e1452.