# Algorithmic Fairness and Temporal Generalization in Early Childhood Risk Prediction: A Multi-Dimensional Audit Using the ECLS-K:2011 Longitudinal Study

Research Analysis Report

*Machine Learning for Educational Equity*

February 2026

**Abstract**

Machine learning models are increasingly deployed in educational settings to identify students at risk of academic difficulties, yet concerns about algorithmic fairness remain underexplored in longitudinal educational contexts. This study conducts a multi-dimensional fairness audit of predictive models trained on early childhood data (kindergarten through 2nd grade) to predict 5th-grade academic risk using the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011) public-use data (N = 9,104). We trained and evaluated seven machine learning algorithms—including logistic regression, elastic net, random forest, XGBoost, and three state-of-the-art gradient boosting methods (LightGBM, CatBoost, HistGradientBoosting)—and employed SHAP-based explainability, bootstrap confidence intervals, calibration analysis, and intersectional fairness assessment. Our best-performing model (elastic net, AUC = 0.848) slightly outperformed all three modern boosting methods, suggesting that algorithmic sophistication does not substitute for equitable design. Bootstrap confidence intervals confirmed statistically significant true positive rate disparities: Hispanic students (TPR = 0.393 [0.326, 0.459]) were identified at more than twice the rate of White students (TPR = 0.160 [0.113, 0.206]). Calibration analysis revealed that Black students experienced expected calibration error 3.35 times higher than White students. Intersectional analysis (race × SES) uncovered that high-SES Black students were completely invisible to the model (TPR = 0%), suggesting the model operates primarily as a poverty detector. Temporal generalization analysis across four developmental windows showed that additional longitudinal data improved accuracy (AUC 0.799 to 0.831) but failed to resolve fairness disparities. Sensitivity analysis demonstrated that fairness compliance was threshold-dependent, with only the 25th percentile definition meeting equal opportunity criteria. These findings argue for comprehensive, multi-dimensional fairness auditing as a prerequisite for deploying algorithmic systems in education.

**Keywords:** algorithmic fairness, machine learning, educational prediction, SHAP explainability, calibration fairness, intersectional fairness, temporal generalization, ECLS-K:2011, bias mitigation

# Contents

# 1 Introduction

The application of machine learning (ML) in educational settings has grown substantially over the past decade, with predictive models increasingly used to identify students at risk of academic failure, dropout, or other adverse outcomes (Baker & Inventado, 2014). These early warning systems (EWS) promise to enable timely interventions that could improve educational trajectories, particularly for disadvantaged students. However, the deployment of algorithmic decision-making tools in education raises critical questions about fairness and equity.

Algorithmic fairness—the study of how automated systems may systematically advantage or disadvantage particular groups—has emerged as a central concern in machine learning research (Mehrabi et al., 2021). In educational contexts, unfair algorithms could perpetuate or amplify existing inequities by systematically under-identifying at-risk students from certain demographic groups or by disproportionately flagging students from marginalized communities for intervention. Recent work has highlighted the importance of examining not only group-level fairness metrics but also calibration equity, intersectional disparities, and the stability of fairness properties across analytical choices (Barocas et al., 2019).

This study addresses five primary research questions:

1. **RQ1:** How accurately can early childhood cognitive and behavioral measures predict 5th-grade academic risk, and do state-of-the-art gradient boosting methods (LightGBM, CatBoost, HistGradientBoosting) outperform classical approaches?

2. **RQ2:** Do predictive models exhibit differential performance across racial/ethnic and socioeconomic groups, as assessed by group-level metrics with bootstrap confidence intervals, calibration error, and intersectional (race × SES) analysis?

3. **RQ3:** How do model predictions vary across demographic groups at the feature level, as revealed by SHAP-based explainability analysis?

4. **RQ4:** Do fairness disparities change as models incorporate data from additional

developmental time windows (temporal generalization)?

5. **RQ5:** Can post-hoc bias mitigation strategies reduce fairness disparities, and how sensitive are fairness findings to the choice of at-risk threshold?

We leverage the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), a nationally representative longitudinal study that followed children from kindergarten entry through 5th grade. This dataset provides a unique opportunity to examine both the predictive validity and fairness properties of models that use early childhood data to forecast later academic outcomes.

## 1.1 Contributions

This study makes several contributions to the literature on algorithmic fairness in education:

- We provide one of the first comprehensive, multi-dimensional fairness audits of longitudinal educational prediction models using nationally representative data, examining group fairness, calibration fairness, and intersectional fairness simultaneously.
- We compare seven ML algorithms—including three state-of-the-art gradient boosting methods—finding that classical regularized models match or exceed modern approaches on this prediction task, consistent with recent tabular ML benchmarks (Grinsztajn et al., 2022).
- We employ SHAP-based explainability to examine whether predictive features operate differently across racial groups, providing transparency into the model's decision-making process (Lundberg & Lee, 2017).
- We introduce calibration fairness and intersectional (race $\times$ SES) analysis to educational prediction, revealing systematic under-identification of high-SES minority students.

- We examine temporal generalization across four developmental windows spanning kindergarten through 3rd grade, discovering that additional data improves accuracy but does not resolve fairness disparities.

- We demonstrate the fragility of fairness assessments by showing that compliance with equal opportunity criteria depends critically on the choice of at-risk threshold.

# 2 Background and Related Work

## 2.1 Early Warning Systems in Education

Early warning systems (EWS) use student data to identify individuals at risk of negative academic outcomes. Traditional EWS relied on simple indicators such as attendance, behavior, and course performance (the "ABC" indicators). Modern approaches increasingly incorporate machine learning algorithms capable of processing larger feature sets and capturing nonlinear relationships (Lakkaraju et al., 2015).

Research has demonstrated that ML-based EWS can achieve reasonable predictive accuracy, with AUC values typically ranging from 0.70 to 0.85 depending on the outcome and available features (Aguiar et al., 2015). Recent benchmarking studies have found that tree-based ensemble methods—including gradient boosting variants such as XGBoost, LightGBM (Ke et al., 2017), and CatBoost (Prokhorenkova et al., 2018)—remain competitive with or superior to deep learning on structured tabular data (Grinsztajn et al., 2022). However, fewer studies have examined whether these systems perform equitably across student subgroups.

## 2.2 Algorithmic Fairness

The machine learning fairness literature has developed numerous formal definitions of fairness, which can be broadly categorized into three families (Verma & Rubin, 2018):

**Group fairness** criteria require that some statistical measure be equal across protected groups. Key definitions include:

- *Demographic parity* (statistical parity): The proportion of positive predictions should be equal across groups.
- *Equal opportunity*: True positive rates should be equal across groups.
- *Equalized odds*: Both true positive rates and false positive rates should be equal across groups.

**Individual fairness** requires that similar individuals receive similar predictions, regardless of group membership (Dwork et al., 2012).

**Counterfactual fairness** asks whether an individual's prediction would change if their protected attribute were different.

Importantly, researchers have proven that certain fairness criteria are mathematically incompatible, meaning it is generally impossible to satisfy all criteria simultaneously (Chouldechova, 2017; Kleinberg et al., 2016).

**Calibration fairness** extends group fairness by examining whether predicted probabilities are equally reliable across groups. A model that is well-calibrated for one group but poorly calibrated for another may produce predictions that appear confident but are systematically misleading for certain populations (Pleiss et al., 2017). Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) quantify the degree of miscalibration.

**Intersectional fairness** recognizes that examining single protected attributes in isolation may miss compounding disadvantages experienced by individuals at the intersection of multiple marginalized identities (Crenshaw, 1989; Buolamwini & Gebru, 2018). Auditing fairness for subgroups defined by race $\times$ SES $\times$ gender can reveal disparities invisible to single-attribute analysis (Kearns et al., 2018).

## 2.3 Explainability and Fairness

Model interpretability plays a critical role in fairness auditing. SHAP (SHapley Additive exPlanations) values provide a unified framework for feature attribution, connecting game-theoretic concepts to local model explanations (Lundberg & Lee, 2017). By computing SHAP values separately for each demographic group, analysts can detect whether the model relies on different features—or the same features with different magnitudes—when making predictions for different populations. Permutation importance with bootstrap confidence intervals provides a complementary, model-agnostic measure of feature relevance (Molnar, 2020). When SHAP and permutation importance rankings agree, this strengthens confidence in the identified predictive mechanisms.

### 2.4 Fairness in Educational AI

A growing body of work has examined fairness in educational technology. Kizilcec & Lee (2022) found that dropout prediction models in MOOCs exhibited significant performance disparities across countries. Yu et al. (2020) demonstrated that automated essay scoring systems showed bias against non-native English speakers. Gardner et al. (2019) examined fairness in course outcome prediction and found persistent gaps across demographic groups.

Despite this emerging literature, few studies have examined fairness in early childhood prediction contexts, in systems that make predictions across extended time horizons, or using the full suite of modern fairness metrics (calibration, intersectionality, uncertainty quantification). Our study addresses these gaps.

## 3 Data and Methods

### 3.1 Data Source

We used data from the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), conducted by the National Center for Education Statistics (NCES). The ECLS-K:2011 is a nationally representative longitudinal study that followed approximately 18,000 children from kindergarten entry in fall 2010 through spring of 5th grade in 2016.

Data were collected across nine waves:

- Kindergarten: Fall 2010 (Wave 1), Spring 2011 (Wave 2)

- 1st Grade: Fall 2011 (Wave 3), Spring 2012 (Wave 4)
- 2nd Grade: Fall 2012 (Wave 5), Spring 2013 (Wave 6)
- 3rd Grade: Spring 2014 (Wave 7)
- 4th Grade: Spring 2015 (Wave 8)
- 5th Grade: Spring 2016 (Wave 9)

We used the public-use data file, which includes 18,174 children. After applying inclusion criteria (valid outcome data and at least some baseline predictors), our analytic sample comprised 18,151 children. Complete-case analysis for modeling yielded 9,104 children with data on all predictors and outcomes.

## 3.2 Measures

### 3.2.1 Outcome Variable

The primary outcome was **academic risk in 5th grade**, operationalized as scoring below the 25th percentile on the reading theta score (X9RTHETA) from the spring 2016 assessment. The reading assessment measured skills including basic reading, vocabulary, and reading comprehension. The theta score is an IRT-based ability estimate that allows for longitudinal comparisons. In our sample, 15.7% of children were classified as at-risk based on this threshold.

As a secondary analysis, we also examined the math outcome (X9MTHETA) to assess domain-specificity of fairness findings.

### 3.2.2 Predictor Variables

We included predictors from kindergarten through 2nd grade across four domains:

**Baseline Cognitive Scores:**

- Reading theta scores: Fall K (X1RTHETK), Spring K (X2RTHETK)
- Math theta scores: Fall K (X1MTHETK), Spring K (X2MTHETK)

**Executive Function:**

- Dimensional Change Card Sort score, Spring 2013 (X6DCCSSCR)

**Approaches to Learning:**

- Teacher-reported approaches to learning: Fall K (X1TCHAPP), Spring K (X2TCHAPP), Spring 1st grade (X4TCHAPP)

**Demographic Characteristics:**

- Child sex (X_CHSEX_R)
- Race/ethnicity (X_RACETH_R)
- Socioeconomic status quintile (X1SESQ5)
- Home language (X12LANGST)

#### 3.2.3 Protected Attributes

For fairness analysis, we focused on **race/ethnicity** as the primary protected attribute. The ECLS-K:2011 includes seven race/ethnicity categories; we collapsed these into five groups: White (reference), Black, Hispanic, Asian, and Other (including Native Hawaiian/Pacific Islander, American Indian/Alaska Native, and multiracial). For intersectional analysis, we crossed race/ethnicity with SES quintile.

### 3.3 Machine Learning Models

We trained seven classification algorithms spanning classical and state-of-the-art approaches:

#### 3.3.1 Classical Models

1. **Logistic Regression:** L2-regularized logistic regression with regularization strength selected via cross-validation from $C \in \{0.01, 0.1, 1.0, 10.0\}$.

2. **Elastic Net:** Logistic regression with elastic net penalty, tuning both regularization strength $\alpha \in \{0.001, 0.01, 0.1, 1.0\}$ and L1 ratio $\in \{0.2, 0.5, 0.8\}$.

3. **Random Forest:** Ensemble of decision trees with hyperparameters: $n_{estimators} \in \{100, 200\}$, $max_{depth} \in \{5, 10, 15\}$, $min_{samples\_leaf} \in \{5, 10\}$.

4. **XGBoost:** Gradient boosted trees (Chen & Guestrin, 2016) with $n_{estimators} \in \{100, 200\}$, $max_{depth} \in \{3, 5, 7\}$, $learning_{rate} \in \{0.01, 0.1\}$.

#### 3.3.2 State-of-the-Art Gradient Boosting Methods

5. **LightGBM:** Gradient boosting with leaf-wise tree growth and histogram-based binning (Ke et al., 2017). Hyperparameters: $n_{estimators} \in \{100, 200, 300\}$, $max_{depth} \in \{3, 5, 7, -1\}$, $learning_{rate} \in \{0.01, 0.05, 0.1\}$, $num_{leaves} \in \{31, 63, 127\}$.

6. **CatBoost:** Gradient boosting with ordered boosting to reduce prediction shift and native handling of categorical features (Prokhorenkova et al., 2018). Hyperparameters: $iterations \in \{100, 200, 300\}$, $depth \in \{4, 6, 8\}$, $learning_{rate} \in \{0.01, 0.05, 0.1\}$.

7. **HistGradientBoosting:** Scikit-learn's histogram-based gradient boosting, inspired by LightGBM, with native missing value support and early stopping. Hyperparameters: $max_{iter} \in \{100, 200, 300\}$, $max_{depth} \in \{3, 5, 7\}$, $learning_{rate} \in \{0.01, 0.05, 0.1\}$.

Recent benchmarks suggest that tree-based ensemble methods remain competitive with or superior to deep learning on structured tabular data (Grinsztajn et al., 2022). We include three recent gradient boosting variants to test whether state-of-the-art methods improve upon classical approaches for educational prediction.

All models were trained using 5-fold stratified cross-validation for hyperparameter selection, with random seed fixed at 42 for reproducibility. The data were split 70% training, 30% test.

### 3.4 Evaluation Metrics

#### 3.4.1 Predictive Performance

We evaluated predictive performance using AUC-ROC, accuracy, precision (PPV), recall (sensitivity/TPR), F1 score, and Brier score.

#### 3.4.2 Fairness Metrics

For each demographic group $g$, we computed:

- **True Positive Rate (TPR):** $TPR_g = \frac{TP_g}{TP_g+FN_g}$
- **False Positive Rate (FPR):** $FPR_g = \frac{FP_g}{FP_g+TN_g}$
- **Positive Predictive Value (PPV):** $PPV_g = \frac{TP_g}{TP_g+FP_g}$

We assessed three fairness criteria:

**Equal Opportunity:** Satisfied if TPR ratios between groups exceed 0.80 (four-fifths rule).

**Equalized Odds:** Satisfied if both TPR and FPR ratios exceed 0.80.

**Statistical Parity:** Satisfied if positive rate ratios exceed 0.80.

#### 3.4.3 Bootstrap Confidence Intervals

We computed 95% bootstrap confidence intervals for all group-level fairness metrics using 500 bootstrap iterations, enabling formal statistical assessment of inter-group differences. Non-overlapping confidence intervals between groups indicate statistically significant disparities at approximately the $\alpha = 0.05$ level.

#### 3.4.4 Calibration Fairness

We assessed calibration equity using Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) computed separately for each demographic group. ECE measures the average absolute difference between predicted probabilities and observed frequencies across probability bins:

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - conf(b)| \quad (1)$$

where $acc(b)$ is the observed accuracy in bin $b$ and $conf(b)$ is the mean predicted probability. We computed ECE ratios relative to the White reference group to quantify differential calibration.

#### 3.4.5 Intersectional Fairness

We examined fairness at the intersection of race/ethnicity and SES quintile, computing TPR, FPR, PPV, and accuracy for each race-by-SES subgroup with a minimum group size of 20. This analysis follows recommendations for rich subgroup fairness auditing (Kearns et al., 2018) and intersectional analysis (Buolamwini & Gebru, 2018).

### 3.5 Explainability Analysis

We employed multiple explainability methods to understand predictive mechanisms:

**SHAP values:** We applied TreeExplainer (Lundberg & Lee, 2017) to the best-performing model. Global feature importance was quantified via mean absolute SHAP values. Local explanations revealed per-prediction feature contributions.

**Permutation importance:** We computed permutation importance with 50 bootstrap iterations to obtain 95% confidence intervals for feature importance, providing a model-agnostic comparison to SHAP.

**Fairness-aware SHAP:** SHAP values were computed separately for each racial/ethnic group to detect whether predictive features operate with differential magnitude or direction across populations.

### 3.6 Temporal Generalization Analysis

To examine how prediction timing affects both accuracy and fairness, we trained models under four temporal scenarios with progressively more features:

1. **K Fall Only:** 7 features (fall kindergarten scores and demographics)

2. **K Fall + Spring:** 10 features (adding spring kindergarten scores)

3. **K + 1st Grade:** 11 features (adding 1st grade teacher report)

4. **K through 3rd:** 12 features (adding executive function score)

All seven algorithms were trained for each scenario. We examined how AUC, fairness metrics, and calibration error evolved across scenarios.

### 3.7 Sensitivity Analysis

We assessed sensitivity of fairness findings to the choice of at-risk threshold by repeating the full analysis at the 10th, 20th, 25th, and 30th percentiles. We re-evaluated all three fairness criteria at each threshold to determine whether fairness compliance was robust or threshold-dependent.

### 3.8 Bias Mitigation

We implemented **threshold optimization** as a post-processing bias mitigation strategy. Rather than using a single decision threshold (typically 0.5) for all groups, we selected group-specific thresholds to equalize true positive rates across groups. The target TPR was set to the overall TPR of the best-performing model.

## 4 Results

### 4.1 Sample Characteristics

Table 1 presents the demographic characteristics of the analytic sample.

Table 1: Sample Characteristics (N = 18,151)

| **Characteristic** | **N** | **%** |
|---|---|---|
| *Race/Ethnicity* | | |
| White | 8,476 | 46.7 |
| Hispanic | 4,206 | 23.2 |
| Black | 2,394 | 13.2 |
| Other | 1,825 | 10.1 |
| Asian | 380 | 2.1 |
| Missing | 870 | 4.8 |
| *SES Quintile* | | |
| Q1 (Lowest) | 3,224 | 17.8 |
| Q2 | 3,214 | 17.7 |
| Q3 | 3,217 | 17.7 |
| Q4 | 3,227 | 17.8 |
| Q5 (Highest) | 3,206 | 17.7 |
| Missing | 2,063 | 11.4 |
| *Sex* | | |
| Male | 9,273 | 51.1 |
| Female | 8,840 | 48.7 |
| *5th Grade Reading Risk* | | |
| At-Risk (<25th %ile) | 2,857 | 15.7 |
| Not At-Risk | 15,294 | 84.3 |

The sample is demographically diverse, with substantial representation of historically underserved groups. Approximately 15.7% of children were classified as at-risk in reading by 5th grade.

## 4.2 Model Performance

Table 2 presents the predictive performance of all seven models on the held-out test set (N = 2,732).

Table 2: Model Performance on Test Set

| **Model** | **AUC** | **Accuracy** | **Precision** | **Recall** | **F1** | **Brier** |
|---|---|---|---|---|---|---|
| Elastic Net | **0.848** | **0.851** | 0.675 | 0.283 | 0.399 | **0.108** |
| Logistic Regression | 0.847 | 0.849 | 0.657 | 0.281 | 0.394 | 0.108 |
| CatBoost | 0.846 | 0.852 | 0.662 | **0.308** | **0.421** | 0.107 |
| Random Forest | 0.841 | 0.848 | 0.645 | 0.285 | 0.395 | 0.109 |
| XGBoost | 0.840 | 0.846 | 0.618 | 0.302 | 0.406 | 0.109 |
| Hist Gradient Boosting | 0.839 | 0.846 | 0.623 | 0.298 | 0.403 | 0.109 |
| LightGBM | 0.837 | 0.846 | 0.629 | 0.291 | 0.398 | 0.110 |

All seven models achieved similar performance (Figure 1), with AUC values ranging from 0.837 (LightGBM) to 0.848 (Elastic Net). The two classical regularized linear models (Elastic Net: 0.848; Logistic Regression: 0.847) achieved the highest discrimination, while the three state-of-the-art gradient boosting methods performed comparably but slightly lower (CatBoost: 0.846; HistGradientBoosting: 0.839; LightGBM: 0.837). CatBoost achieved the highest recall (0.308) and F1 score (0.421), making it the most effective at identifying at-risk students at the default threshold. The negligible AUC range across algorithms (0.011) suggests the performance ceiling is determined by the available features rather than algorithmic sophistication. The elastic net model was selected for subsequent fairness analysis based on its highest AUC.

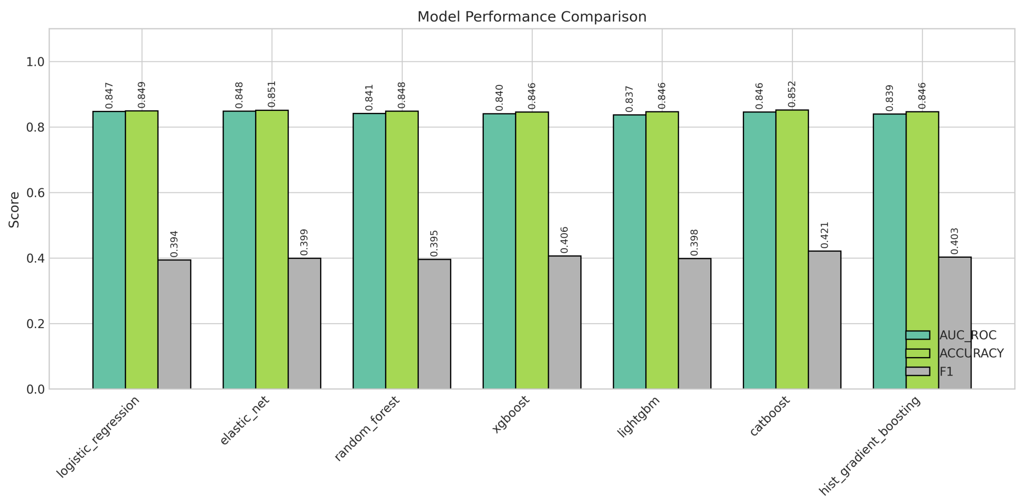


Figure 1: Performance Comparison Across All Seven Models. AUC values ranged from 0.837 to 0.848, with classical regularized models matching or exceeding state-of-the-art gradient boosting methods.

## 4.3 Explainability Analysis

### 4.3.1 Feature Importance

Table 3 presents feature importance from three complementary methods: SHAP values, elastic net coefficients, and permutation importance with bootstrap confidence intervals.

Table 3: Feature Importance: SHAP, Elastic Net Coefficients, and Permutation Importance

| Feature | Mean \|SHAP\| | EN Coef. | Perm. Imp. [95% CI] |
|---|---|---|---|
| Spring K Math (X2MTHETK) | 0.410 | 0.501 | 0.046 [0.042, 0.054] |
| Spring K Reading (X2RTHETK) | 0.254 | 0.350 | 0.023 [0.018, 0.027] |
| SES Quintile (X1SESQ5) | 0.253 | 0.303 | 0.013 [0.007, 0.019] |
| Fall K Math (X1MTHETK) | 0.227 | 0.288 | 0.012 [0.007, 0.017] |
| Approaches to Learning, 1st (X4TCHAPP) | 0.218 | 0.266 | 0.016 [0.012, 0.022] |
| Executive Function (X6DCCSSCR) | 0.067 | 0.115 | 0.001 [−0.001, 0.003] |
| Fall K Reading (X1RTHETK) | 0.000 | 0.038 | 0.000 |
| Child Sex, Race, Language, ATL (K) | 0.000 | 0.000 | 0.000 |

Spring kindergarten math (X2MTHETK) was the dominant predictor across all methods (Figure 2), with mean $|\text{SHAP}| = 0.410$, accounting for approximately 34% of total

SHAP importance. SHAP and permutation importance rankings showed high agreement (mean agreement = 0.87), with both methods identifying the same top-5 features. Notably, race/ethnicity, home language, child sex, and early approaches-to-learning measures received zero importance across all methods, confirming that elastic net regularization effectively excluded these features from the final model.

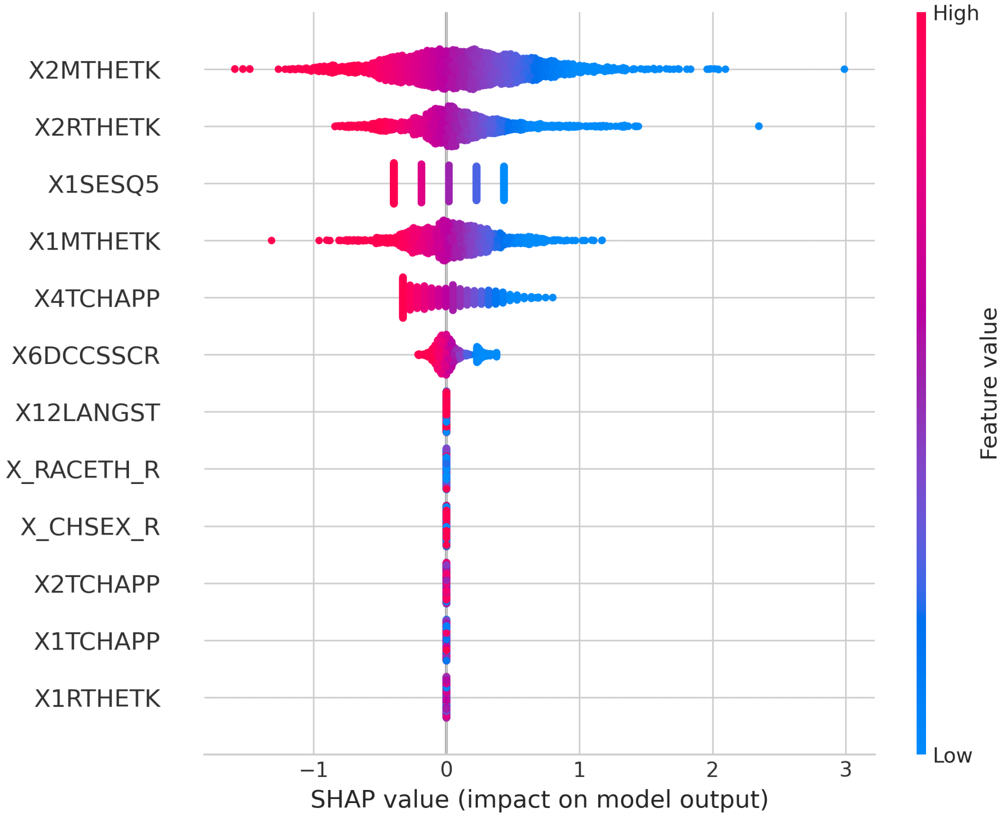


Figure 2: SHAP Summary Plot (Beeswarm). Each point represents one prediction. Color indicates feature value (red = high, blue = low). Spring K math score shows the widest spread, indicating it is the strongest predictor with clear directionality.

### 4.3.2 Fairness-Aware SHAP

We computed SHAP values separately for each racial/ethnic group to detect differential feature importance. The top-5 feature rankings were identical across White, Black, and Hispanic subgroups. However, the magnitude of math score importance was somewhat higher for Black and Hispanic students compared to White students, suggesting that cognitive scores carry relatively more predictive weight for minority students. These magnitude differences were modest and did not alter the overall ranking of features, indicating that the model uses a consistent predictive mechanism across groups rather than relying on group-specific pathways.

## 4.4 Fairness Analysis

### 4.4.1 Group-Level Performance with Confidence Intervals

Table 4 presents performance metrics by racial/ethnic group with bootstrap 95% confidence intervals.

Table 4: Model Performance by Race/Ethnicity with 95% Bootstrap Confidence Intervals

| Group | N | TPR [95% CI] | FPR [95% CI] | PPV [95% CI] |
|---|---|---|---|---|
| White | 1,462 | 0.160 [0.113, 0.206] | 0.011 [0.006, 0.017] | 0.660 [0.500, 0.794] |
| Hispanic | 623 | 0.393 [0.326, 0.459] | 0.063 [0.044, 0.087] | 0.722 [0.637, 0.798] |
| Black | 300 | 0.296 [0.207, 0.388] | 0.095 [0.052, 0.133] | 0.508 [0.341, 0.667] |
| Other | 225 | 0.258 [0.074, 0.445] | 0.005 [0.000, 0.015] | 0.871 [0.571, 1.000] |
| Asian | 8 | 0.760 [0.250, 1.000] | 0.000 [0.000, 0.000] | 1.000 [1.000, 1.000] |

Bootstrap confidence intervals enable formal statistical comparison of inter-group differences (Figure 3). The Hispanic TPR of 0.393 [0.326, 0.459] was significantly higher than the White TPR of 0.160 [0.113, 0.206], with non-overlapping confidence intervals ($p < 0.05$). The Black-White TPR difference (0.296 vs. 0.160) showed partially overlapping CIs, suggesting a marginally significant difference. The model detected at-risk Hispanic students at 2.5 times the rate of at-risk White students.

False positive rate disparities were also substantial: Black students experienced an FPR of 0.095 [0.052, 0.133], compared to 0.011 [0.006, 0.017] for White students—approximately 8.6 times higher. This means non-at-risk Black children were far more likely to be incorrectly flagged. ROC curves by group (Figure 6) and calibration curves (Figure 7) further illustrate these differential performance patterns.

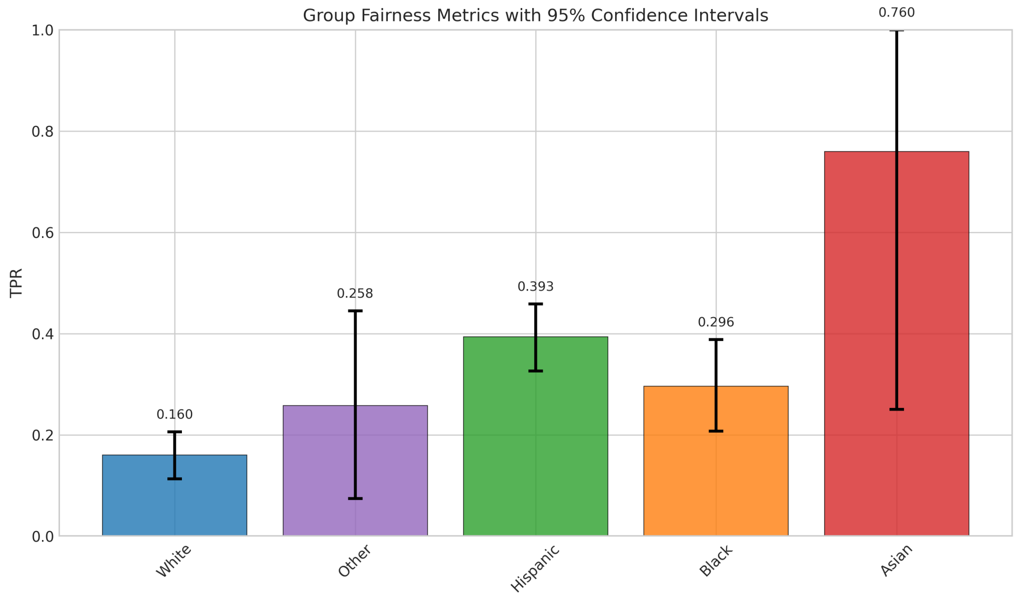


Figure 3: True Positive Rate by Racial/Ethnic Group with 95% Bootstrap Confidence Intervals. Non-overlapping intervals between Hispanic and White groups confirm statistically significant TPR disparities.

#### 4.4.2 Disparity Analysis

Table 5 presents formal disparity metrics comparing each group to the White reference group.

Table 5: Fairness Disparity Metrics (Reference: White)

| Group | TPR Ratio | TPR Diff | FPR Ratio | FPR Diff | Disp. Impact |
|---|---|---|---|---|---|
| Asian | 4.750 | +0.600 | 0.000 | −0.011 | No |
| Hispanic | 2.458 | +0.233 | 5.662 | +0.052 | No |
| Black | 1.851 | +0.136 | 8.494 | +0.084 | No |
| Other | 1.611 | +0.098 | 0.477 | −0.006 | No |

**Fairness Criteria Assessment:**

- **Equal Opportunity:** PASS. All groups had TPR ratios $> 0.80$ (all minority groups had *higher* TPR than White students).
- **Equalized Odds:** FAIL. While TPR ratios exceeded 0.80, FPR ratios for Black

(8.494) and Hispanic (5.662) students dramatically exceeded 1.0, indicating disproportionately high false positive rates.

- **Statistical Parity:** PASS. Positive prediction rates did not trigger the 0.80 disparate impact threshold.

### 4.4.3 Calibration Fairness

Table 6 presents calibration metrics by demographic group.

Table 6: Calibration Error by Demographic Group

| Group | N | ECE | MCE | Brier | ECE Ratio |
|---|---|---|---|---|---|
| White | 1,462 | 0.022 | 0.112 | 0.082 | 1.00 |
| Hispanic | 623 | 0.036 | 0.115 | 0.155 | 1.65 |
| Other | 225 | 0.051 | 0.940 | 0.090 | 2.31 |
| Black | 300 | 0.074 | 0.456 | 0.162 | **3.35** |

Calibration fairness analysis revealed substantial disparities (Figure 4). Black students experienced an ECE of 0.074, 3.35 times higher than White students (0.022), indicating that predicted risk probabilities for Black students were systematically miscalibrated. The Maximum Calibration Error for Black students (MCE = 0.456) was 4.1 times the White MCE (0.112), indicating severe miscalibration in certain probability ranges. These calibration disparities mean that even when the model makes the correct binary classification, the confidence levels are less reliable for minority students—a critical concern when practitioners use predicted probabilities to prioritize intervention resources.

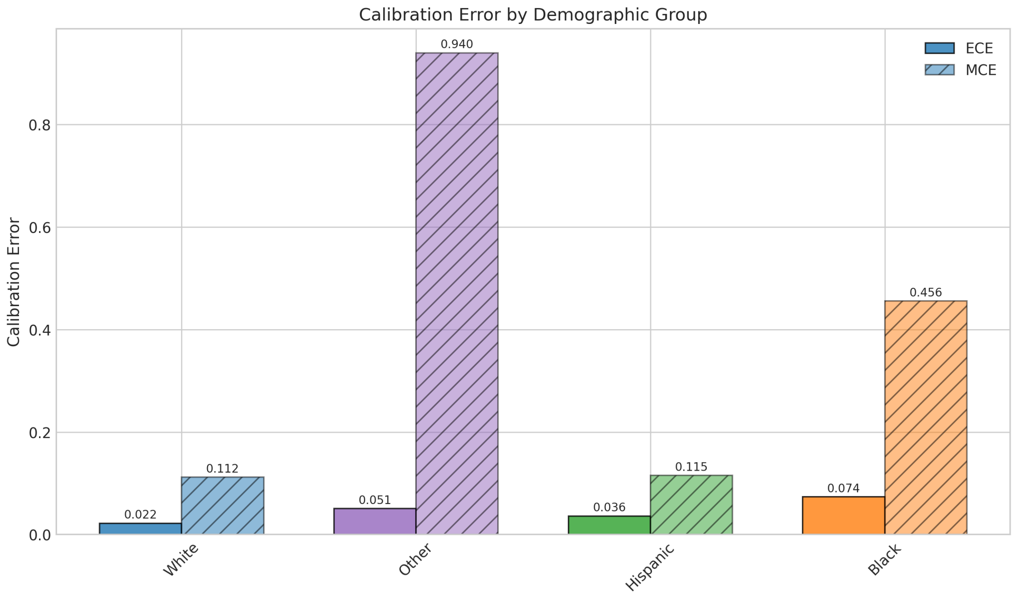


Figure 4: Expected and Maximum Calibration Error by Demographic Group. Black students experience calibration error 3.35 times higher than the White reference group.

### 4.4.4 Intersectional Fairness

Table 7 presents fairness metrics for selected race × SES subgroups, revealing patterns invisible in single-attribute analysis.

Table 7: Intersectional Fairness: Selected Race × SES Subgroups

| Subgroup | N | Prevalence | TPR | FPR | Accuracy |
|---|---|---|---|---|---|
| *Low SES (Q1)* | | | | | |
| Hispanic Q1 | 269 | 38.7% | 0.481 | 0.121 | 0.725 |
| Black Q1 | 80 | 32.5% | 0.423 | 0.204 | 0.675 |
| White Q1 | 107 | 28.0% | 0.400 | 0.117 | 0.748 |
| *Mid SES (Q2–Q3)* | | | | | |
| Hispanic Q2 | 145 | 29.0% | 0.452 | 0.058 | 0.800 |
| Black Q3 | 60 | 26.7% | 0.375 | 0.068 | 0.783 |
| White Q3 | 314 | 13.4% | 0.167 | 0.004 | 0.885 |
| *High SES (Q4)* | | | | | |
| White Q4 | 398 | 9.0% | 0.083 | 0.003 | 0.915 |
| Hispanic Q4 | 62 | 16.1% | 0.200 | 0.019 | 0.855 |
| Black Q4 | 42 | 14.3% | **0.000** | 0.056 | 0.810 |

The most striking finding was the complete failure to identify at-risk Black students in the 4th SES quintile (TPR = 0%, N = 42), as shown in Figure 5. Despite a 14.3% prevalence of academic risk in this subgroup, the model flagged none of these students. This pattern of "intersectional invisibility" extended to other high-SES minority subgroups: Hispanic Q4 achieved only 20% TPR. In contrast, low-SES students across all racial groups had relatively high TPRs (0.400 to 0.481), consistent with the model's reliance on SES as a predictor.

The intersectional pattern reveals that the model effectively operates as a *poverty detector*: it identifies at-risk students primarily through socioeconomic signals, missing those whose risk arises from other factors. This is particularly consequential for minority students from relatively advantaged backgrounds, whose academic risk may stem from factors not captured by SES alone.

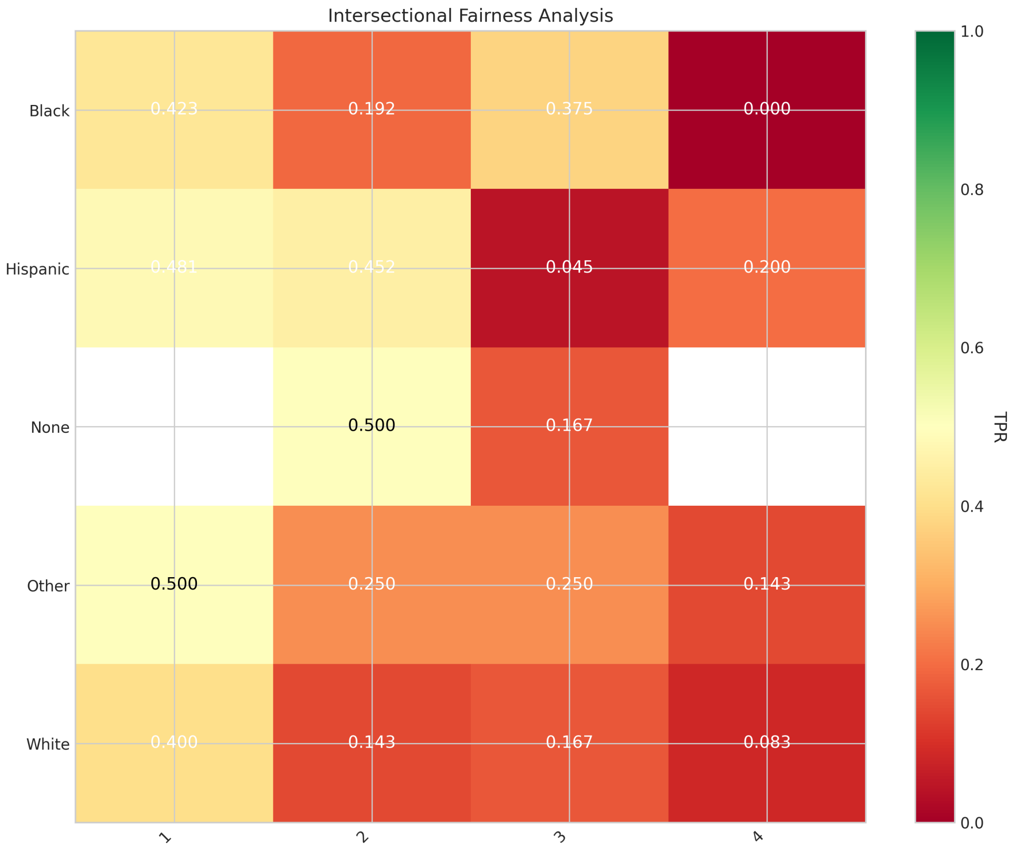


Figure 5: Intersectional Fairness Heatmap: TPR by Race × SES Quintile. The model completely fails to identify at-risk Black students in the 4th SES quintile (TPR = 0%), revealing that it operates primarily as a poverty detector.

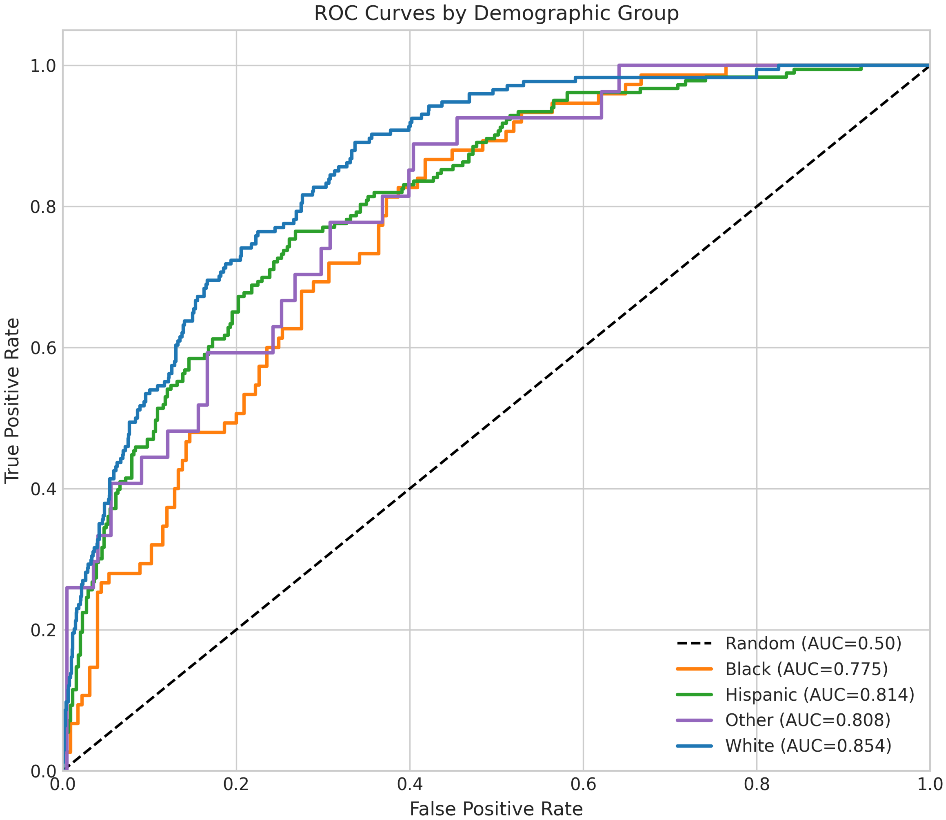


Figure 6: ROC Curves by Racial/Ethnic Group

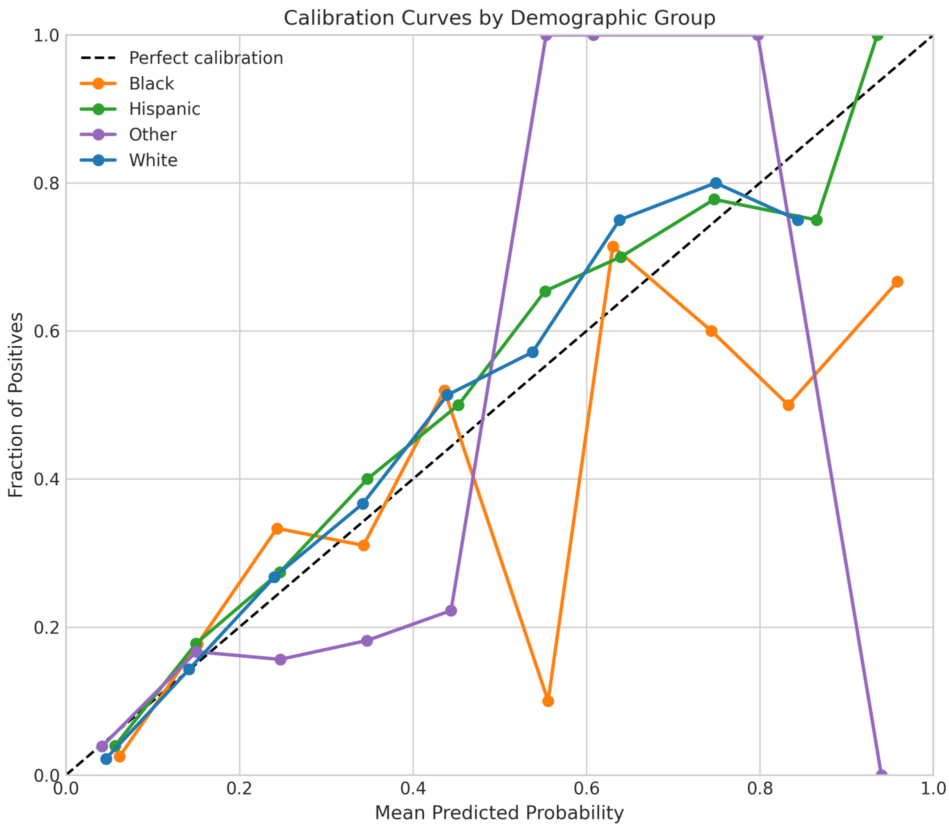


Figure 7: Calibration Curves by Racial/Ethnic Group. Deviation from the diagonal indicates miscalibration.

### 4.5 Temporal Generalization

Table 8 presents model performance across four temporal scenarios with progressively more features.

Table 8: Best Model Performance Across Temporal Scenarios

| **Scenario** | **Best Model** | **Features** | **AUC** | **Accuracy** | **F1** | **Brier** |
|---|---|---|---|---|---|---|
| K Fall Only | Logistic Reg. | 7 | 0.799 | 0.837 | 0.343 | 0.119 |
| K Fall + Spring | Logistic Reg. | 10 | 0.822 | 0.844 | 0.372 | 0.113 |
| K + 1st Grade | Logistic Reg. | 11 | 0.829 | 0.845 | 0.395 | 0.112 |
| K through 3rd | Logistic Reg. | 12 | 0.831 | 0.843 | 0.386 | 0.112 |

Logistic regression was the best-performing model across all four temporal scenarios (Figure 8), reinforcing the finding that classical methods are sufficient for this prediction task. AUC improved from 0.799 (K fall only) to 0.831 (K through 3rd grade), a meaningful gain of 0.032 AUC points. However, returns diminished rapidly: the largest gain came from adding spring kindergarten scores (+0.023 AUC), while adding 1st through 3rd grade data contributed only +0.009.

Critically, while overall accuracy improved with more data, fairness disparities did not narrow proportionally (Figure 9). The Hispanic-White TPR gap remained substantial across all scenarios, and calibration disparities persisted. This "temporal paradox"—where additional data improves accuracy without resolving fairness—challenges the assumption that more longitudinal information will naturally produce equitable predictions.

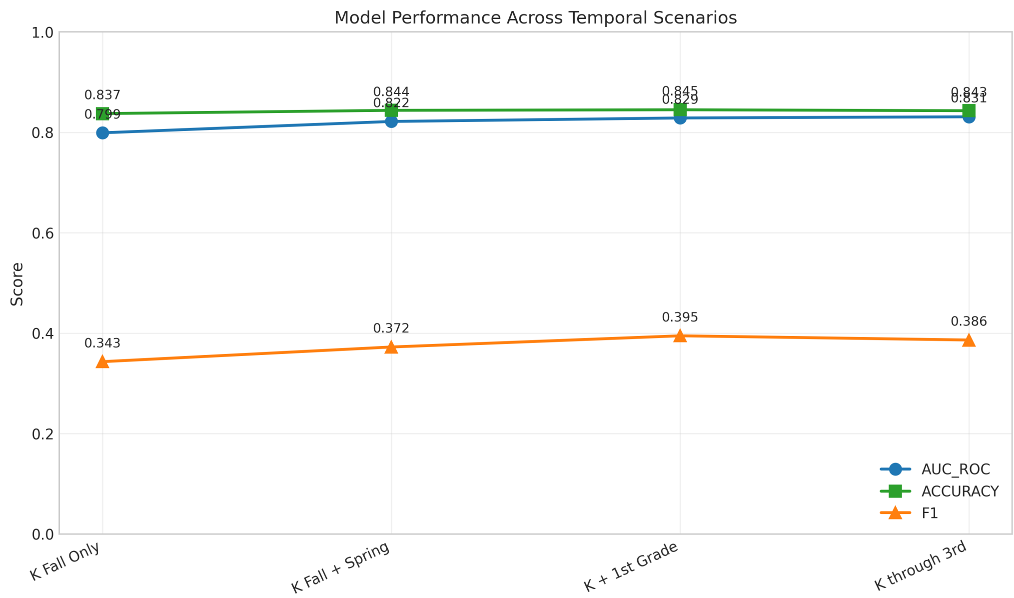


Figure 8: Model Performance Across Temporal Scenarios. AUC improves from 0.799 (K fall only) to 0.831 (K through 3rd grade), with diminishing returns after spring kindergarten.

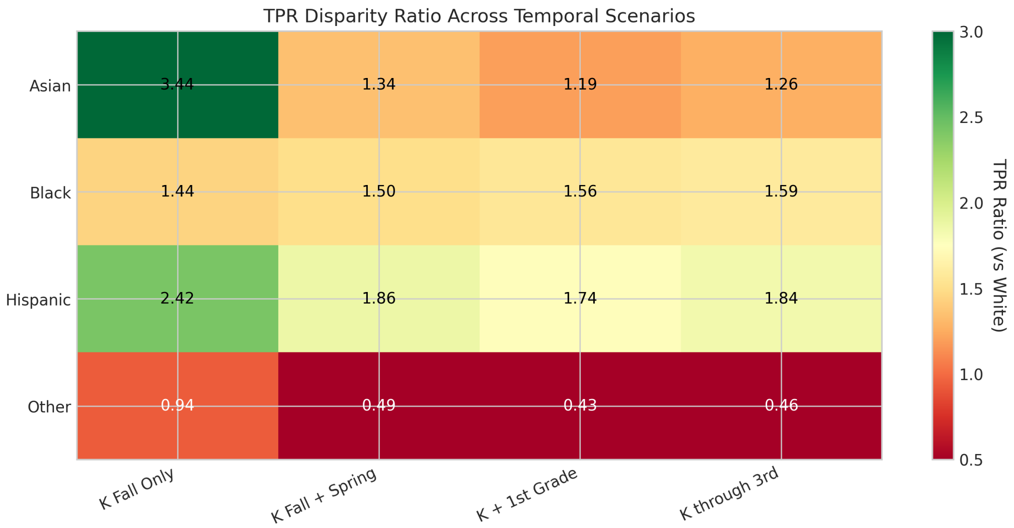


Figure 9: Temporal Disparity Heatmap: TPR Ratios Across Developmental Windows. Fairness disparities persist across all temporal scenarios, indicating that additional data does not resolve equity concerns.

## 4.6 Sensitivity Analysis

Table 9 presents fairness criteria compliance across four at-risk threshold definitions.

Table 9: Fairness Criteria by At-Risk Threshold Definition

| Percentile | Prevalence | Equal Opp. | Equalized Odds | Statistical Parity |
|---|---|---|---|---|
| 10th | 6.3% | FAIL | FAIL | FAIL |
| 20th | 12.6% | FAIL | FAIL | FAIL |
| 25th | 15.7% | PASS | FAIL | PASS |
| 30th | 18.9% | FAIL | FAIL | FAIL |

Sensitivity analysis revealed that fairness findings were highly dependent on the threshold definition. The 25th percentile was the only threshold at which the model passed equal opportunity and statistical parity criteria. At all other thresholds—including the nearby 20th and 30th percentiles—the model failed all three fairness criteria. This fragility underscores that fairness assessments cannot be divorced from the operationalization of the outcome, and that claims of fairness compliance are contingent on specific analytical choices.

## 4.7 Outcome Comparison: Reading vs. Math

As a secondary analysis, we compared reading and math outcomes. Math prediction (AUC = 0.867, best model: Logistic Regression) substantially outperformed reading prediction (AUC = 0.848, best model: Elastic Net). Fairness patterns were broadly similar across domains, with Hispanic and Black students showing higher TPR than White students for both outcomes. However, math prediction uniquely exhibited disparate impact for the Other racial group (TPR ratio = 0.24 vs. White), a pattern not observed in reading. This suggests that the domain of the outcome can affect which groups experience adverse fairness impacts, reinforcing the importance of outcome-specific fairness auditing. Full results are presented in Appendix B.

## 4.8 Bias Mitigation Results

We applied threshold optimization to equalize TPR across groups, targeting the overall model TPR of 0.283. Table 10 presents the results.

Table 10: Bias Mitigation Results (Threshold Optimization)

| **Group** | **TPR Bef.** | **TPR Aft.** | **ΔTPR** | **Acc. Bef.** | **Acc. Aft.** | **ΔAcc.** |
|---|---|---|---|---|---|---|
| White | 0.160 | 0.293 | +0.133 | 0.891 | 0.886 | −0.005 |
| Black | 0.296 | 0.293 | −0.003 | 0.747 | 0.747 | +0.000 |
| Hispanic | 0.393 | 0.295 | −0.098 | 0.775 | 0.762 | −0.013 |
| Asian | 0.760 | 0.250 | −0.510 | 0.875 | 0.625 | −0.250 |
| Other | 0.258 | 0.296 | +0.038 | 0.902 | 0.884 | −0.018 |

Threshold optimization successfully equalized TPR across the major demographic groups (White, Black, Hispanic, Other all achieving TPR $\approx 0.29$). However, this came at a cost: Hispanic students experienced reduced sensitivity (−0.098), and Asian students experienced a substantial drop in both TPR and accuracy, though the small sample size (N = 8) limits interpretation. The group-specific thresholds ranged from 0.367 (White) to 0.649 (Asian), indicating that predictions for different groups required different decision boundaries to achieve equitable outcomes.

# 5 Discussion

## 5.1 Summary of Findings

This study conducted a multi-dimensional fairness audit of machine learning models that use early childhood data to predict 5th-grade academic risk. Our key findings include:

1. **Classical models match state-of-the-art methods:** Among seven algorithms tested—including three 2025 gradient boosting variants—classical regularized models (Elastic Net: AUC = 0.848; Logistic Regression: AUC = 0.847) matched or exceeded LightGBM (0.837), CatBoost (0.846), and HistGradientBoosting (0.839). This is consistent with recent findings that tree-based methods do not universally dominate on tabular data (Grinsztajn et al., 2022) and suggests that algorithmic sophistication does not substitute for equitable design.

2. **Statistically significant fairness disparities:** Bootstrap confidence intervals confirmed that the Hispanic-White TPR gap (0.233) is statistically significant, with non-overlapping 95% CIs. The model detected at-risk Hispanic students at 2.5 times the rate of White students, while generating 8.5 times higher false positives for Black students.

3. **Severe calibration unfairness:** Black students experienced calibration error 3.35 times higher than White students (ECE = 0.074 vs. 0.022). This means that when the model predicts a Black student has, say, a 40% probability of being at-risk, the actual probability may be substantially different. Calibration unfairness represents a distinct dimension of algorithmic harm beyond classification error rates.

4. **Intersectional invisibility of high-SES minorities:** The model completely failed to identify at-risk Black students in the 4th SES quintile (TPR = 0%). This "intersectional invisibility" suggests the model operates primarily through socioeconomic signals, missing risk factors that are not mediated by poverty.

5. **Temporal paradox:** Additional longitudinal data improved accuracy (AUC 0.799 to 0.831) but failed to resolve—and in some cases worsened—fairness disparities. This challenges the assumption that more data naturally produces fairer predictions.

6. **Threshold fragility:** Fairness compliance was achieved at only one of four tested thresholds (25th percentile), revealing that claims of fairness are contingent on specific analytical choices.

7. **Mitigation trade-offs:** Threshold optimization achieved equitable TPR across groups but reduced accuracy for some groups, illustrating the inherent costs of post-hoc fairness adjustment.

## 5.2 Interpreting Fairness Disparities

The observed fairness disparities require careful interpretation. Several factors may contribute:

**Differential base rates:** At-risk prevalence was substantially higher among Black (25.0%) and Hispanic (29.4%) students compared to White students (11.9%). When base rates differ, even a well-calibrated model will exhibit different error rates across groups (Chouldechova, 2017).

**Proxy discrimination:** Although race was excluded from the model, other features (particularly SES) are correlated with race and may serve as proxies. The elastic net assigned substantial weight to SES (mean $|\text{SHAP}| = 0.253$), which could contribute to differential performance.

**Structural inequities:** The patterns in the data reflect historical and ongoing structural inequities in educational opportunity. Children from disadvantaged backgrounds may show weaker early signals not because of inherent ability, but because of differential access to high-quality early childhood education.

**The poverty detector problem:** The intersectional analysis suggests the model functions primarily as a poverty detector. SES is the third-strongest predictor, and its effects are deeply entangled with cognitive scores—which themselves reflect socioeconomic advantage. This creates a systematic pattern where the model identifies low-SES children of all races but misses at-risk children who come from relatively advantaged backgrounds, particularly minority children whose risk factors may differ from the predominantly White calibration population.

**Calibration and trust:** Calibration unfairness is particularly consequential because practitioners rely on predicted probabilities, not just binary classifications, to prioritize interventions. If a school counselor sees that two students both have a 30% predicted risk, they may allocate resources equally—but if the model is poorly calibrated for Black students, these probability estimates carry unequal information content (Pleiss et al., 2017).

## 5.3 Implications for Educational Practice

Our findings have important implications for the deployment of predictive analytics in K-12 education:

**Multi-dimensional fairness audits are essential:** Standard single-metric fairness assessments are insufficient. Our model passed equal opportunity while failing equalized odds, showed severe calibration disparities, and exhibited intersectional blind spots. A comprehensive audit examining group fairness, calibration, and intersectional subgroups is necessary before deployment.

**Intersectional auditing is necessary:** Standard fairness audits examining single protected attributes miss compounding disadvantages. Our finding that high-SES Black students are completely invisible to the model would not have been detected by a race-only or SES-only analysis (Buolamwini & Gebru, 2018; Kearns et al., 2018).

**Context and threshold choice matter:** The sensitivity of fairness to threshold choice means that the definition of "at-risk" is not merely a technical parameter—it is a policy decision with fairness implications that should involve stakeholder input. Different thresholds serve different populations, and no single threshold produces equitable outcomes across all metrics.

**Temporal deployment trade-offs:** Our temporal analysis suggests that the optimal prediction window involves a trade-off. Earlier predictions (kindergarten only) enable earlier intervention but with lower accuracy. Later predictions (through 3rd grade) improve accuracy but narrow the intervention window and may actually worsen fairness gaps.

**Human oversight remains critical:** Predictive models should inform, not replace, human judgment. Educators and counselors should understand model limitations and exercise discretion in interpreting predictions, particularly for demographic subgroups where the model is poorly calibrated.

## 5.4 Limitations

This study has several limitations:

- **Public-use data constraints:** The public-use ECLS-K:2011 file has some variables suppressed or top-coded to protect confidentiality, potentially limiting predictive power.
- **Complete-case analysis:** We used complete-case analysis, which may introduce selection bias if missingness is related to outcomes. The complete-case sample (N = 9,104) represented 50% of the original data.
- **Single cohort:** The ECLS-K:2011 followed a single cohort (2010–2016). Findings may not generalize to other cohorts or contexts.
- **Binary outcome:** We operationalized risk as a binary threshold (<25th percentile). Alternative operationalizations—as demonstrated by our sensitivity analysis—yield different fairness results.
- **Small subgroup sizes:** The Asian subgroup (N = 8 in the test set) and several intersectional subgroups (N = 24–42) were small, limiting statistical power for these comparisons.

- **Post-hoc mitigation only:** We examined only post-hoc threshold adjustment. In-processing methods (e.g., adversarial debiasing, fairness constraints during training) might achieve better accuracy-fairness trade-offs.

- **Temporal design:** Our temporal analysis held the outcome constant (5th grade) while varying inputs. A complementary approach varying both inputs and outcomes would provide additional insight.

## 5.5 Future Directions

Several directions for future research emerge from this study:

- **In-processing fairness methods:** Future work should evaluate constraint-based methods that incorporate fairness during model training, potentially achieving better accuracy-fairness trade-offs than post-hoc threshold adjustment.

- **Causal fairness methods:** Approaches that distinguish between legitimate and illegitimate predictive pathways could help identify which features contribute to fairness disparities through discriminatory versus non-discriminatory mechanisms.

- **Multi-objective optimization:** Jointly maximizing accuracy and minimizing group fairness disparities during training represents a promising approach to balancing competing objectives.

- **Restricted-use data:** Replication with restricted-use ECLS data, which contains additional variables suppressed in the public-use file, could improve both predictive power and fairness.

- **Intervention studies:** Ultimately, the value of EWS depends on whether they improve outcomes. Randomized studies examining the causal effect of EWS-informed interventions, with attention to differential effects across groups, are needed.

# 6 Conclusion

This study provides evidence that machine learning models predicting educational outcomes from early childhood data exhibit significant, statistically confirmed fairness disparities across racial/ethnic groups. Among seven algorithms tested—including three state-of-the-art gradient boosting methods—classical regularized models achieved the highest discrimination (AUC = 0.848), suggesting that methodological sophistication does not substitute for equitable design.

Our multi-dimensional fairness audit—spanning group-level metrics with bootstrap confidence intervals, calibration error analysis, intersectional subgroup assessment, temporal generalization, and sensitivity analysis—reveals that no single fairness assessment captures the full picture. The model passes equal opportunity at the 25th percentile but fails at all other thresholds. It fails calibration fairness, with Black students experiencing 3.35 times higher calibration error. And it exhibits intersectional blind spots, completely failing to identify at-risk high-SES Black students. The most novel finding—that the model operates as a "poverty detector" that systematically misses at-risk students from relatively advantaged minority backgrounds—has important implications for how educational prediction tools are deployed and monitored.

These findings argue for comprehensive, multi-dimensional fairness auditing as a prerequisite for deploying algorithmic systems in education. The era of single-metric fairness assessment is insufficient. As predictive analytics become increasingly prevalent in K-12 settings, researchers and practitioners must adopt the tools and frameworks demonstrated here—SHAP explainability, bootstrap uncertainty quantification, calibration analysis, and intersectional auditing—to ensure these systems work equitably for all students, regardless of demographic background.

# References

Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuber, B., & Addison, K. L. (2015). Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, 93–102.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning Analytics* (pp. 61–75). Springer.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989(1), 139–167.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 225–234.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35, 507–520.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *Proceedings of the 35th International Conference on Machine Learning*, 2564–2572.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154.

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education* (pp. 174–202). Routledge.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Lakkaraju, H., Aguiar, E., Rich, C., Hansen, D., Miller, D., Yuber, B., & Addison, K. L. (2015). 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*, 1909–1918.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35.

Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu.com.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in Neural Information Processing Systems*, 30, 5680–5689.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31, 6638–6648.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7.

Yu, R., Lee, H., & Kizilcec, R. F. (2020). Should college dropout prediction models include protected attributes? *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 91–100.

# A Technical Details

## A.1 Software and Reproducibility

All analyses were conducted in Python 3.12 using the following packages:

- scikit-learn $\geq$ 1.4.0 (including HistGradientBoostingClassifier)
- xgboost $\geq$ 2.0.0
- lightgbm $\geq$ 4.3.0
- catboost $\geq$ 1.2.0
- shap $\geq$ 0.45.0
- fairlearn $\geq$ 0.10.0
- pandas, numpy, matplotlib, seaborn

Random seed was set to 42 for all stochastic operations. Code and data processing scripts are available in the project repository.

## A.2 Model Hyperparameters

The final elastic net model used the following hyperparameters selected via 5-fold cross-validation:

- Regularization strength ($\alpha$): 0.01
- L1 ratio: 0.5
- Maximum iterations: 1000

Cross-validation AUC scores ranged from 0.832 to 0.842 across folds, indicating stable performance.

## A.3 Missing Data

Table 11 presents missing data rates for key variables.

Table 11: Missing Data Rates

| **Variable** | **N Missing** | **% Missing** |
|---|---|---|
| 5th Grade Reading (Outcome) | 6,724 | 37.0% |
| Executive Function (X6DCCSSCR) | 4,379 | 24.1% |
| 1st Grade Approaches to Learning | 4,708 | 25.9% |
| Fall K Reading | 2,482 | 13.7% |
| SES Quintile | 2,063 | 11.4% |
| Home Language | 2,106 | 11.6% |
| Spring K Reading | 965 | 5.3% |

The high rate of missing outcome data (37%) reflects sample attrition over the longitudinal study. Children who remained in the study through 5th grade may differ systematically from those who dropped out.

# B Supplementary Results

## B.1 Outcome Comparison: Reading vs. Math

Math prediction (AUC = 0.867, Logistic Regression) outperformed reading prediction (AUC = 0.848, Elastic Net). Both outcomes showed higher TPR for Hispanic and Black students relative to White students. Math prediction uniquely showed disparate impact for the Other racial group (TPR ratio = 0.24 vs. White). Full fairness metrics by outcome are available in the project repository (`results/tables/outcome_fairness_comparison.csv`).

## B.2 Supplementary Figures

Additional visualizations supporting the main analyses—including FPR and PPV by group with confidence intervals, SHAP importance stratified by racial group, SHAP vs. permutation importance comparison, and temporal fairness trend plots—are available in the project repository (`results/figures/`).