Predicting Two-Year Public Versus Four-Year Postsecondary Enrollment from Ninth-Grade Measures and the High School Transcript: Calibration and Subgroup Fairness in HSLS:09

EDM-ARS

Automated Research System

Author Note

Note: This manuscript was generated by EDM-ARS, an automated educational data mining research system, under human oversight. **Disclosures statement:** The authors declare no competing interests. **Funding:** No external funding was received.

Abstract

Among students who enroll in postsecondary education, the sector of enrollment—two-year public versus four-year—carries substantial consequences for cost, completion, and transfer pathways, yet most predictive models target any enrollment rather than sector sorting. Using HSLS:09 data for 12,942 students who ever attended a postsecondary institution, we trained five machine learning models to predict two-year public enrollment from ninth-grade measures and the high school transcript record. The best model (XGBoost) achieved an AUC of 0.704 (95% clustered CI [0.675, 0.730]), statistically indistinguishable from logistic regression (AUC = 0.699). Academic GPA dominated predictions (SHAP mean |value| = 0.517), followed by SES and ninth-grade math achievement. The model was well calibrated overall (Brier = 0.208, ECE = 0.031). Subgroup performance varied materially: AUC ranged from 0.669 to 0.781 across race/ethnicity groups and from 0.627 to 0.731 across SES quintiles. These results suggest that sector-specific prediction is feasible but moderate in accuracy, and that subgroup-specific calibration and performance monitoring are essential before deploying such models in advising contexts.

Keywords: postsecondary sector, machine learning, calibration, subgroup fairness, HSLS:09, two-year college

Predicting Two-Year Public Versus Four-Year Postsecondary Enrollment from Ninth-Grade Measures and the High School Transcript: Calibration and Subgroup Fairness in HSLS:09

Introduction

Postsecondary advising in the United States increasingly operates on the output of predictive models. Early-warning systems flag students at risk of not enrolling in college, and advising tools route students toward resources based on predicted trajectories (He et al., 2025). The implicit assumption behind most of these systems is that the relevant prediction target is *whether* a student enrolls in postsecondary education. But enrollment is not a single destination. The two-year public sector enrolls roughly a third of all undergraduates and serves a disproportionate share of students from lower-income families, students of color, and first-generation students (Yao et al., 2025). A student who enrolls at a community college faces a different cost structure, a different set of transfer pathways, and different completion odds than a student who enrolls at a four-year institution (Khilnaney et al., 2025). A model that predicts "any college enrollment" without distinguishing sector therefore answers a question that is, for many advising decisions, the wrong question.

The distinction matters for practice. Community colleges are geographically dispersed, low-cost, and open-access; four-year institutions are selective, expensive, and concentrated in particular regions (Means, 2024). The factors that push a student toward one sector rather than the other are not identical to the factors that predict enrollment at all. Cost sensitivity, proximity to home, information about transfer articulation, and the intensity of high school academic preparation all operate differently on the two-year versus four-year margin than on the enrolled versus not-enrolled margin (Plasman & Myles, 2024). An early-warning system that cannot distinguish these margins will misallocate advising effort: it may flag a student as "on track" because she is predicted to enroll somewhere, while missing that she is predicted to enroll in a sector with a much lower

probability of bachelor's degree completion.

The research literature reflects this gap. A substantial body of work applies machine learning to postsecondary enrollment prediction, and the retrieved corpus for this study contains eighteen papers in the college-enrollment-by-prediction-ML cell alone. The dominant finding across this literature is that high school grade point average, socioeconomic status, and prior standardized achievement are the strongest predictors of any college enrollment (Guanin-Fajardo et al., 2024; He et al., 2025). But the same corpus contains no study that models the two-year-public versus four-year margin specifically, conditional on postsecondary entry, using the High School Longitudinal Study of 2009 (HSLS:09) with ninth-grade measures plus the high school transcript record. The nearest retrieved paper predicts any college enrollment among low-SES HSLS:09 students and reports that high school GPA, parental educational expectations, and peer plans are the top predictors (He et al., 2025). That study does not distinguish sector. The sector-sorting question remains open.

This study addresses that gap. We ask: among HSLS:09 students who ever attended a postsecondary institution, how accurately can enrollment at a two-year public institution rather than a four-year institution be predicted from ninth-grade measures and the high school transcript record, how well calibrated are those predictions, and is predictive accuracy consistent across sex, race/ethnicity, and socioeconomic groups?

The conditioning on postsecondary entry is a deliberate design choice, not a convenience. The analytic sample comprises students who ever attended a postsecondary institution, which means the model estimates sector sorting *among enrollees*, not sector sorting among all high school students. This is the population for which the question is practically meaningful: advising systems that operate on enrolled students need to know, conditional on enrollment, where a student is likely to land. The selection condition also has a substantive consequence that we state here and return to in the Limitations: findings generalize to the subpopulation of postsecondary entrants, not to all ninth graders.

We make three contributions. First, we provide the first sector-specific prediction model for the two-year-public versus four-year margin using HSLS:09, with ninth-grade measures and the high school transcript record as predictors. Second, we assess calibration explicitly, reporting Brier score, expected calibration error, and calibration slope alongside discrimination. Calibration is underreported in educational prediction work (Tiruneh et al., 2025; Xia & Chen, 2026), and a model that discriminates well but is miscalibrated can mislead advisors about the actual probability of two-year enrollment. Third, we conduct a subgroup fairness analysis across sex, race/ethnicity, and socioeconomic status, reporting subgroup-specific AUCs and flagging gaps larger than five percentage points. Prior work on fairness in educational prediction has documented that aggregate accuracy can mask substantial subgroup disparities (Opoku et al., 2025; Raftopoulos et al., 2025), and we treat the subgroup analysis as a first-class result rather than an appendix.

The findings are, in brief, these. The best model achieves an AUC of approximately 0.70, which is moderate: the model can rank students by predicted probability of two-year public enrollment, but it is not a precise sorting instrument. Academic GPA dominates the predictor set, with a SHAP importance more than 2.5 times that of the next single predictor, socioeconomic status. The model is reasonably well calibrated overall, with an expected calibration error of 0.031. Subgroup performance, however, varies materially: the model is more accurate for Asian students and higher-SES students than for White students and lower-SES students, with subgroup AUC gaps exceeding ten percentage points. These gaps are the paper's central substantive finding, and they carry direct implications for how such models should be deployed in advising contexts.

The remainder of the paper proceeds as follows. We review the literature on postsecondary enrollment prediction, sector sorting, and calibration and fairness in educational machine learning. We then describe the HSLS:09 data, the analytic sample, the predictor set, and the modeling and evaluation procedures. We report results for model comparison, feature importance, calibration, subgroup performance, school-level variance,

and a moderation analysis of the GPA-sector relationship by SES. We close with a discussion of what the model can and cannot do for advising, the fairness implications of the subgroup gaps, and specific limitations and future directions.

Related Work

Predicting Postsecondary Enrollment with Machine Learning

The use of machine learning to predict postsecondary outcomes has matured into a substantial literature, with a dense cluster of work addressing college enrollment specifically. Across these studies, a consistent set of predictors emerges: high school grade point average, socioeconomic status, and prior standardized achievement dominate feature-importance rankings, while model performance typically settles in a moderate range. He et al. (2025) analyzed HSLS:09 data for 5,223 low-SES ninth graders and found that random forest achieved the highest classification accuracy at 67.73%, with overall high school GPA, parental educational expectations, and peer college-going plans as the top predictors of any college enrollment. That study is representative of the field's orientation: the outcome is binary enrollment, the population is often restricted by socioeconomic status, and the substantive question is who enrolls at all rather than where.

The HSLS:09 dataset has supported several such investigations because it links rich ninth-grade survey data to high school transcripts and postsecondary records. Plasman and Myles (2024) used HSLS:09 to examine how engineering career and technical education credits relate to four-year enrollment among students with learning disabilities, finding that each E-CTE credit was associated with a higher likelihood of four-year rather than sub-baccalaureate enrollment. Zambrano and Baker (2024) took a different angle, using middle school digital learning platform data to show that mastery of specific mathematics topics predicted later college enrollment and STEM career selection. These studies share a common limitation for our purposes: they treat postsecondary destination as a binary or coarse categorical outcome and do not isolate the two-year-public versus four-year contrast within the enrolled population.

Beyond HSLS:09, the broader prediction literature reinforces the same pattern. Guanin-Fajardo et al. (2024) applied CRISP-DM methodology to a dataset of 6,690 college students and found XGBoost achieved an AUC of 87.75% for academic success prediction, while Tang et al. (2024) compared ten algorithms for student persistence and reported that random forest with SMOTE outperformed logistic regression. Guevara-Reyes et al. (2025) trained XGBoost and random forest on 50,000 student records, reporting an R^2 of 0.91 for academic performance prediction with socioeconomic conditions and infrastructure among the top features. Mahawar and Rattan (2024) and Sharma and Mansotra (2026) similarly pursued ensemble and neural approaches for student outcome prediction, with the latter reporting 96.20% accuracy on the OULAD dataset. What unites these studies is their focus on a single aggregate outcome: enrollment, persistence, or performance. None of them models the sector of enrollment conditional on having enrolled, which is the question that motivates the present study.

The methodological toolkit in this literature is also relatively uniform. Logistic regression, random forest, XGBoost, and occasionally neural networks are the standard model families, with AUC or accuracy as the primary metric. Tiruneh et al. (2025) demonstrated in a clinical context that logistic regression and XGBoost often achieve statistically indistinguishable discrimination, a finding that anticipates our own model-comparison result. Wang et al. (2025) showed that feature selection can reduce dimensionality without sacrificing accuracy, while Guazzo et al. (2024) reported that some prediction tasks are simply difficult regardless of model complexity. These methodological observations matter for our study because they suggest that the moderate AUC we report for sector sorting may reflect the intrinsic difficulty of the task rather than a deficiency in modeling approach.

Two-Year versus Four-Year Sorting: What Is Known

The substantive literature on community college versus four-year enrollment is extensive and theoretically rich, though it rarely intersects with the machine learning

prediction literature. Sociological and economic research has established that sector choice is driven by a distinct set of mechanisms from the decision to enroll at all. Cost sensitivity is paramount: two-year public institutions offer substantially lower tuition, making them attractive to students from lower-income families and those averse to debt (Means, 2024). Proximity matters as well; community colleges are geographically dispersed, and students often choose them because they can live at home and maintain family and work obligations. Information constraints also play a role: students whose parents did not attend college, or whose high schools lack strong college-going cultures, may not learn about four-year options or financial aid pathways in time to act on them.

Curricular intensity in high school is a well-documented predictor of sector sorting. Students who complete advanced mathematics and science coursework are more likely to enroll in four-year institutions, while those with weaker academic preparation disproportionately enter two-year colleges (Plasman & Myles, 2024). This is not merely a selection effect; the transcript record itself signals readiness for four-year work, and admissions processes at four-year institutions use it as a gatekeeping mechanism. Schudde et al. (2025) showed that ever-English-learner status predicts college type among Texas public college students, with ever-ELs more likely to enter community colleges, while Khilnaney et al. (2025) documented the mentoring and support structures that two-year STEM students rely on. These studies confirm that sector sorting is patterned by academic preparation, language background, and institutional context in ways that are theoretically distinct from the any-enrollment decision.

The distinction between any-enrollment and sector-sorting prediction is not merely academic. A model that predicts whether a student will enroll in any postsecondary institution answers a different question than one that predicts, conditional on enrollment, whether that student will attend a two-year public or a four-year institution. The former is relevant for dropout prevention and college-going culture interventions; the latter is relevant for advising students about transfer pathways, for understanding equity in access

to bachelor's degree programs, and for institutions planning articulation agreements. Yao et al. (2025) made a related point in the context of retention prediction at community colleges, arguing that models trained at four-year institutions do not transfer cleanly to two-year contexts because the student populations and institutional missions differ. This insight extends to our study: a model trained to predict any enrollment would not necessarily capture the mechanisms that sort enrolled students into sectors.

Calibration and Fairness in Educational Prediction

A parallel literature has emerged on the calibration and fairness of educational prediction models, driven by recognition that discrimination metrics like AUC are insufficient for deployment decisions. Calibration, which measures whether predicted probabilities match observed frequencies, is critical when predictions inform individual-level decisions such as advising or intervention targeting. Carriero et al. (2024) demonstrated through simulation that class imbalance corrections can harm calibration even when discrimination is unaffected, while Yang et al. (2024) found empirically that random oversampling and undersampling generally do not improve AUC and can produce miscalibrated risk estimates. These findings are directly relevant to our study, which reports calibration metrics alongside discrimination and finds that the model is well calibrated overall despite the moderate AUC.

Fairness in educational prediction has received increasing attention, with a growing body of work documenting subgroup performance gaps. Opoku et al. (2025) systematically examined accuracy-fairness trade-offs in student performance prediction on the OULAD dataset, finding that standard models exhibit bias and that mitigation techniques can reduce disparities with modest accuracy costs. Gao and Ni (2025) introduced FairEduNet, an adversarial framework that reduces dependence on sensitive attributes while maintaining accuracy, and Raftopoulos et al. (2025) compared preprocessing, in-processing, and post-processing bias mitigation techniques across educational datasets. Raftopoulos et al. (2024) applied fairness metrics to admission prediction, while Lünich and Keller

(2024) found that decision tree simplicity, not accuracy, drove student perceptions of fairness. These studies establish that subgroup performance gaps are common in educational ML and that they carry ethical weight, but they rarely report calibration alongside fairness, and none addresses sector-sorting prediction specifically.

The clinical prediction literature offers methodological lessons that transfer to education. Brown et al. (2025) compared large language models to locally trained ML for clinical prediction, reporting both AUROC and Brier scores and finding that traditional ML substantially outperformed LLMs on both discrimination and calibration. Davis et al. (2025) documented fairness drift over time in clinical models, showing that subgroup gaps can emerge post-deployment even when initial assessments are favorable. Broden et al. (2026) integrated conformal prediction with multiverse analysis to audit group-conditional coverage, arguing that uncertainty quantification must be part of fairness evaluation. These methodological advances have not yet been fully absorbed into educational prediction research, where calibration is often unreported and subgroup analysis is frequently limited to aggregate accuracy comparisons.

Gap Analysis

The retrieved literature, taken together, leaves a clear gap. The college enrollment prediction literature is dense but oriented toward binary any-enrollment outcomes, with HSLS:09 studies like He et al. (2025) predicting whether low-SES students enroll at all rather than where they enroll. The substantive literature on two-year versus four-year sorting is theoretically rich but rarely employs machine learning methods or reports predictive performance metrics. The fairness and calibration literature has developed sophisticated methods for auditing subgroup performance and probability calibration, but has not applied them to the sector-sorting question. No retrieved paper models two-year-public versus four-year enrollment conditional on postsecondary entry using HSLS:09 with ninth-grade measures plus the high school transcript record, and none reports calibration alongside subgroup fairness for this outcome. The present study fills

that gap by combining the sector-specific outcome definition with a rigorous evaluation framework that includes clustered bootstrap confidence intervals, calibration metrics, and subgroup performance analysis across race/ethnicity and socioeconomic status.

Methods

Data and Sample

We used the High School Longitudinal Study of 2009 (HSLS:09) public-use file, a nationally representative study that followed a cohort of ninth-grade students through postsecondary education. The analytic sample was restricted to students who ever attended a postsecondary institution, defined as a valid response of "Yes" on the postsecondary attendance indicator (X4EVRATNDCLG). Among these students, we further restricted to those with a usable value on the outcome variable, X4EVR2YPUB, which records whether the student's first postsecondary institution was a two-year public institution. This yielded an analytic sample of 12,942 students, of whom 39.0% attended a two-year public institution and 61.0% attended a four-year institution. Conditioning on postsecondary entry is a deliberate design choice: the research question concerns sector sorting among students who enrolled in postsecondary education, not the prediction of whether any enrollment occurred. Findings therefore generalize to the population of HSLS:09 students who entered postsecondary education, not to all ninth graders.

The train/test split was school-aware. Because school identifiers are suppressed in the public-use file, we reconstructed pseudo-school clusters by grouping students with matching school-level variable profiles (school climate scale, counselor perception scales, school control, locale, and region). This yielded 927 pseudo-school clusters against an expected 944 based on the HSLS:09 sampling frame, with a mean cluster size of 14.0 students (median 13, range 1–53). We then applied StratifiedGroupKFold so that no school appeared in both the training and test partitions. The training set contained 10,364 students from 738 clusters; the test set contained 2,578 students from 189 clusters. The headline test metric is therefore computed over students in schools never seen during

training, which provides a cross-context generalization estimate. Because the reconstruction cannot be verified against true school membership, claims of generalization to unseen schools are provisional on the reconstruction approximating true school membership.

Predictors and Missing Data

We used 12 raw predictors drawn from two waves: ninth-grade measures (base year) and the high school transcript record (second follow-up). The transcript predictors were academic GPA across academic courses (X3TGPAACAD), total credits earned (X3TCREDTOT), English credits (X3TCREDENG), mathematics credits (X3TCREDMAT), and highest science course level (X3THISCI). The ninth-grade predictors were standardized mathematics achievement (X1TXMTSCOR), socioeconomic status composite (X1SES), school urbanicity (X1LOCALE), school control (X1CONTROL), census region (X1REGION), sex (X1SEX), and race/ethnicity (X1RACE). All predictors were measured before the outcome, satisfying temporal ordering. After one-hot encoding of categorical variables, the model matrix contained 28 columns.

Missing data were handled by predictor type. Continuous transcript variables (GPA, total credits, English credits, mathematics credits) had 4.2–4.5% missingness and were imputed with the median. The ninth-grade math score and SES composite had 7.3% missingness and were imputed with IterativeImputer, a multivariate imputation method that uses all other predictors. Categorical variables (science course level, locale, control, region, sex, race/ethnicity) had low missingness (0–4.3%) and were handled by complete-case analysis within the encoding step. We assumed missingness was missing at random (MAR) conditional on observed predictors. The outcome variable has structural missingness in the full HSLS:09 file (26.7% of all students lack a valid X4EVR2YPUB record); this is a population restriction rather than random dropout, and the analytic sample represents students with a valid sector record.

Models and Evaluation

We trained five model families: logistic regression, random forest, XGBoost, elastic net, and a stacking ensemble that combined the individual models via a logistic meta-learner. Logistic regression and elastic net are linear models; random forest and XGBoost are tree-based ensembles; the stacking ensemble aggregates all base models. All models were trained on the school-aware training partition and evaluated on the held-out test partition. The primary metric was area under the receiver operating characteristic curve (AUC), which measures the probability that a randomly selected two-year enrollee receives a higher predicted probability than a randomly selected four-year enrollee. We also report accuracy, precision, recall, and F1 for completeness.

Because students are nested within schools, standard row-level bootstrap confidence intervals underestimate uncertainty for predictors that vary at the school level. We computed the intraclass correlation (ICC) of the outcome on the test set using the reconstructed school clusters, and we computed both standard and clustered bootstrap confidence intervals for the best model's AUC. The clustered bootstrap resamples whole schools with replacement, preserving within-school correlation. We report both intervals so the difference is auditable.

We assessed calibration with four metrics computed on the test set: Brier score (mean squared error of predicted probabilities), expected calibration error (ECE, the weighted average absolute difference between predicted probability and observed frequency across decile bins), calibration slope, and calibration intercept. A calibration slope of 1 and intercept of 0 indicate perfect calibration; slope below 1 indicates overconfidence at the extremes.

We conducted subgroup fairness analysis for the three protected attributes specified in the research question: sex (X1SEX), race/ethnicity (X1RACE), and socioeconomic status (X1SES). For each attribute, we computed the AUC separately for each group on the test set. For the continuous SES composite, we grouped students into quintile bins. We

flagged any attribute whose maximum minus minimum subgroup AUC exceeded 5 percentage points as a fairness concern. We also conducted a moderation analysis testing whether the relationship between academic GPA and sector enrollment varied by SES, using a likelihood-ratio test for the interaction term and a bootstrap confidence interval for the difference in incremental AUC between the top and bottom SES tertiles.

Two methodological limitations are acknowledged here and elaborated in the Limitations section. First, we did not model school-level random effects; the clustered bootstrap accounts for within-school correlation in the primary metric's uncertainty, but a full mixed-effects model would require the restricted-use file with true school identifiers. Second, we did not apply HSLS:09 survey weights. The machine learning estimators used here do not support complex survey variance estimation, and applying weights without proper variance estimation would produce correctly weighted point estimates with incorrect standard errors. The reported metrics reflect unweighted sample performance.

This study was conducted using EDM-ARS, an automated educational data mining research system. All data preparation, model training, evaluation, and manuscript generation were performed programmatically.

Results

Model Comparison: All Models Perform Similarly, and None Is Strongly Predictive

Table 1 reports the held-out test performance of the five model families. The primary metric is the area under the receiver operating characteristic curve (AUC), computed on the school-aware test partition of 2,578 students drawn from 189 pseudo-school clusters not seen during training. XGBoost achieved the highest point estimate (AUC = 0.704, 95% CI [0.683, 0.723]), followed closely by the stacking ensemble (AUC = 0.705, 95% CI [0.685, 0.723]), random forest (AUC = 0.699, 95% CI [0.679, 0.717]), logistic regression (AUC = 0.699, 95% CI [0.678, 0.717]), and elastic net (AUC = 0.692, 95% CI [0.671, 0.710]). The clustered bootstrap confidence interval for XGBoost,

which accounts for within-school correlation, is [0.675, 0.730], slightly wider than the standard interval, consistent with the moderate intraclass correlation reported below.

The paired cluster-bootstrap comparison between XGBoost and the logistic regression baseline yielded an AUC difference of 0.0055 (95% CI [−0.0040, 0.0147]), which does not exclude zero. The models are therefore statistically indistinguishable on the primary metric. Logistic regression captures essentially all of the predictive signal available in these predictors, and the more flexible tree-based and ensemble methods add no reliable discrimination. This is a substantive finding rather than a methodological disappointment: the relationship between ninth-grade measures, the transcript record, and sector sorting is approximately linear in the log-odds, and no interaction structure of the kind tree ensembles can exploit is present at a magnitude the data can resolve. For interpretability and deployment, the simpler linear model is preferable.

Academic GPA Dominates Sector Prediction, with SES and Ninth-Grade Math Achievement as Secondary Signals

Figure 1 displays the SHAP summary plot for the best individual model (XGBoost), and Figure 2 shows the mean absolute SHAP values. Because several predictors are dummy-encoded, the per-dummy SHAP signs are interpreted relative to each variable's reference category; the grouped view in Table 2 aggregates dummy columns to their parent construct and is the appropriate basis for substantive interpretation.

Academic GPA (X3TGPAACAD) dominates the grouped ranking with a mean absolute SHAP value of 0.517, more than 2.5 times the next single predictor. The direction is negative: lower academic GPA pushes predictions toward two-year public enrollment. The next most important groups are highest science course taken (0.212), socioeconomic status (0.186), census region (0.180), ninth-grade mathematics achievement (0.130), and school control (0.122). Total mathematics credits (0.059), race/ethnicity (0.056), total credits (0.043), English credits (0.033), and sex (0.013) contribute comparatively little.

The dominance of academic GPA is educationally coherent. The transcript record,

and especially the grade point average computed across academic courses, is the single strongest signal of the curricular trajectory that routes students toward one sector or the other. Students with lower GPAs are systematically more likely to enroll at two-year public institutions, whether through self-selection, counselor steering, or admission constraints at four-year institutions. Socioeconomic status and ninth-grade mathematics achievement operate as secondary signals in the same direction: lower SES and lower early math achievement both push predictions toward the two-year sector. The partial dependence plots in Figures 3, 4, and 5 confirm these monotonic relationships.

The Model Is Well Calibrated Overall

Calibration was assessed on the test partition using the best individual model (XGBoost). The Brier score is 0.208, and the expected calibration error (ECE) across ten probability bins is 0.031. The calibration slope is 0.842 with an intercept of -0.185. Figure 6 displays the calibration curve.

These metrics indicate that the model is well calibrated on average: predicted probabilities track observed frequencies closely, with an ECE of 0.031 meaning that, across the ten bins, the average absolute deviation between predicted and observed event rates is about three percentage points. The calibration slope below 1 and the negative intercept together indicate slight overconfidence at the extremes: the model assigns somewhat more extreme probabilities than the data warrant. The deviation is modest and would not, by itself, disqualify the model from use as a screening signal. The Brier score of 0.208 is close to the value expected for a model with AUC near 0.70 on an outcome with roughly 39% prevalence, confirming that the discrimination and calibration metrics are internally consistent.

Subgroup Performance Varies Materially by Race/Ethnicity and SES

Table 3 reports subgroup AUCs for race/ethnicity and socioeconomic status. The research question also specified sex as a subgroup attribute, but sex was not available in the protected-attributes file for the test partition, so sex-specific performance could not be

computed; this gap is addressed in the Limitations section.

For race/ethnicity, the model is most accurate for Asian students (AUC = 0.778, $n = 226$) and least accurate for students identifying with more than one race (AUC = 0.669, $n = 227$) and White students (AUC = 0.688, $n = 1{,}374$). Black students (AUC = 0.709, $n = 244$) and Hispanic students with race specified (AUC = 0.700, $n = 353$) fall between. The American Indian/Alaska Native cell reports AUC = 0.781 but rests on only 12 students and is too small for reliable estimation; it should not be interpreted as evidence of high accuracy for that group. The Native Hawaiian/Pacific Islander cell ($n = 9$) was excluded entirely. The range across the interpretable cells is 0.669 to 0.778, a gap of 0.112.

For socioeconomic status, the model is most accurate in the highest SES quintile (AUC = 0.731, $n = 474$) and least accurate in the lowest quintile (AUC = 0.627, $n = 475$). The middle three quintiles fall between (0.674 to 0.680). The gap between the highest and lowest quintiles is 0.104. This gradient is monotonic in the expected direction: the model discriminates sector sorting better for higher-SES students than for lower-SES students.

School-Level Variance Is Substantial

The intraclass correlation coefficient for the outcome on the test partition is 0.205, which falls in the moderate range. Approximately one-fifth of the variance in two-year public enrollment is attributable to school-level differences, consistent with the school-aware split and the decision to report clustered bootstrap confidence intervals. This is both a methodological caution and a substantive finding: school context matters for sector sorting in ways that student-level predictors alone do not capture. The clustered confidence interval for the best model ([0.675, 0.730]) is wider than the standard interval ([0.683, 0.723]), as expected when ICC is non-negligible.

Moderation Analysis: The GPA–Sector Relationship Varies by SES

A likelihood-ratio test for the interaction between academic GPA and socioeconomic status in predicting two-year public enrollment yielded $\chi^2(1) = 30.16$, $p < 0.001$. The incremental AUC attributable to the GPA block within SES tertiles was 0.0042 higher in

the top tertile than in the bottom tertile (95% CI [0.0011, 0.0080]). The interaction is statistically significant, but the magnitude is small: GPA's predictive value for sector sorting is slightly stronger among higher-SES students. This result is consistent with the subgroup gradient reported above and reinforces the conclusion that a single model serves higher-SES students somewhat better than lower-SES students, though the difference is modest in absolute terms.

Discussion

What the Model Can and Cannot Do for Advising

The headline result is a moderate but real predictive signal: the best model distinguishes students who enrolled at two-year public institutions from those who enrolled at four-year institutions with an AUC of 0.704 (95% clustered CI [0.675, 0.730]). In practical terms, this means that if we draw one student at random from each sector, the model assigns a higher predicted probability of two-year enrollment to the two-year enrollee about 70% of the time. That is useful as a screening signal, but it is not a precise sorting instrument. A substantial share of two-year enrollees receive predicted probabilities below the sample mean, and a substantial share of four-year enrollees receive predicted probabilities above it. Any advising system built on these predictions would therefore need to treat the output as a probabilistic flag rather than a deterministic assignment.

The model comparison results reinforce this caution. XGBoost, the best-performing individual model, was not statistically distinguishable from logistic regression (paired cluster-bootstrap difference in AUC = 0.0055, 95% CI [-0.0040, 0.0147]). The stacking ensemble performed essentially identically (AUC = 0.705). This is a substantive finding in its own right: the predictive signal in these ninth-grade and transcript measures is largely linear and additive, and a simple, interpretable logistic regression captures nearly all of it. For deployment contexts where transparency matters—and educational advising is one such context—the simpler model is the better default. This pattern echoes findings in clinical prediction, where logistic regression frequently matches or exceeds more complex learners

on tabular data with modest sample sizes (Brown et al., 2025; Tiruneh et al., 2025).

The Dominance of Academic GPA and the Role of SES

The SHAP analysis identified academic GPA (X3TGPAACAD) as the dominant predictor, with a mean absolute SHAP value of 0.517—more than 2.5 times the next single predictor, socioeconomic status (X1SES, 0.186). When dummy-coded science course level is grouped at the parent-variable level, it emerges as the second most important construct (0.212), followed by SES, region, ninth-grade math achievement, and school control. The direction of these effects is educationally coherent: lower GPA, lower SES, and lower ninth-grade math achievement all push predictions toward two-year public enrollment. This is consistent with the broader literature on any-enrollment prediction, where high school GPA and SES are consistently the strongest predictors (Glandorf et al., 2024; He et al., 2025). What is notable here is that the same hierarchy holds for sector sorting conditional on enrollment. The transcript record, and academic GPA in particular, is not merely a signal of whether a student goes to college; it is also a signal of where they go.

The moderation analysis adds a subtle but statistically detectable layer to this picture. The interaction between GPA and SES was significant (LRT = 30.16, $p < 0.001$), and the incremental AUC attributable to the GPA block was slightly larger in the highest SES tertile than in the lowest (difference = 0.0042, 95% CI [0.0011, 0.0080]). The magnitude is small, and we do not want to overstate it. But the direction is consistent with a substantive interpretation: among higher-SES students, academic preparation is a more decisive determinant of sector choice, perhaps because lower-SES students face constraints—cost, proximity, information—that attenuate the link between preparation and sector. This is a hypothesis for future work, not a conclusion of the present study, but it aligns with qualitative evidence that postsecondary opportunity for lower-SES and rural students is shaped by intersecting financial, spatial, and informational constraints (Means, 2024).

Fairness Implications of Subgroup Performance Gaps

The subgroup analysis is the most consequential finding for practice. The model's accuracy varies materially across race/ethnicity and SES. For race/ethnicity, AUC ranged from 0.669 for students identifying as more than one race (non-Hispanic) to 0.778 for Asian students (non-Hispanic), a gap of 0.112. For SES quintiles, AUC ranged from 0.627 in the lowest quintile to 0.731 in the highest, a gap of 0.104. Both gaps exceed the 5-percentage-point threshold we set as a fairness concern.

Two caveats are necessary before interpreting these numbers. First, the American Indian/Alaska Native cell contains only 12 students, and its AUC of 0.781 is not a reliable estimate; the race gap should be read primarily through the larger cells. Second, the sex subgroup analysis could not be computed because the protected-attribute file did not contain the sex variable, so the fairness claim in our research question is only partially addressed. We state this plainly rather than implying completeness.

With those caveats, the pattern is clear and educationally meaningful. The model is more accurate for Asian students and for higher-SES students, and less accurate for White students, students identifying as more than one race, and lower-SES students. This is not a property of any particular algorithm; it reflects the structure of the data and the outcome. If such a model were deployed in an advising context, the differential accuracy would translate into differential flagging: lower-SES students would be misclassified more often, and because the errors would not be random, the system could reinforce the very disparities it was designed to address. This concern is well documented in the fairness literature, which shows that aggregate accuracy can mask substantial subgroup-level degradation (Alsayed, 2026; Davis et al., 2025; Opoku et al., 2025). The practical implication is direct: any deployment of a sector-prediction model should report subgroup-specific calibration and performance, and should not assume that a single threshold serves all groups equally (Yao et al., 2025).

Implications for Research and Practice

For research, the study makes a case that sector-specific prediction is a distinct task from any-enrollment prediction. The two tasks share predictors—GPA, SES, prior achievement—but they answer different questions, and the answer to the sector question is more modest. Future work should model sector choice with more granular postsecondary data: institution-level characteristics, distance to the nearest two-year college, articulation agreements, and state-level tuition policy. These variables are not in the HSLS:09 public-use file, but they are precisely the factors that the substantive literature identifies as drivers of two-year versus four-year sorting (Plasman & Myles, 2024; Schudde et al., 2025).

For practice, the study's clearest message is about reporting standards. Advising and early-warning systems increasingly report overall AUC, but overall AUC is not enough. The calibration results here are reassuring—Brier score 0.208, expected calibration error 0.031, calibration slope 0.842—but calibration on the full sample does not guarantee calibration within subgroups. The subgroup AUC gaps we report are exactly the kind of information that a practitioner needs before trusting a model with real students. We recommend that any deployment of a sector-prediction model include subgroup-specific performance monitoring as a standing requirement, not a one-time audit (Broden et al., 2026; Davis et al., 2025). The model can flag students who are likely to enroll in two-year public institutions, and that flag has value for targeting information about transfer pathways and articulation. But the flag is only as fair as the model that produces it, and fairness here is not a property of the algorithm alone—it is a property of the data, the outcome, and the deployment context.

Limitations and Future Directions

The most consequential limitation of this study is the multilevel structure of the HSLS:09 data and our incomplete account of it. HSLS:09 sampled approximately 25 students within each of 944 schools, creating a nested structure in which students from the same school share unobserved contextual influences. School identifiers (`SCH_ID`) are

suppressed in the public-use file, but school-level variables such as school climate, counselor perceptions, school control, locale, and region are identical for all students within the same school. We reconstructed pseudo-school clusters by grouping students with matching school-level variable profiles, recovering 927 clusters against an expected 944 (a ratio of 0.98), with a mean of 14.0 students per cluster (median 13.0, range 1–53). The intraclass correlation for the outcome was 0.205, a moderate value indicating that roughly one-fifth of the variance in two-year public enrollment is attributable to school-level differences. We therefore computed clustered bootstrap confidence intervals for the primary metric, which were wider than the standard intervals (clustered CI [0.675, 0.730] versus standard CI [0.683, 0.723]), confirming that ignoring clustering would understate uncertainty. However, this approach provides clustered confidence intervals but does not estimate school-level random effects. A full mixed-effects model would require either the restricted-use file with true `SCH_ID` or adaptation of the pipeline to use `statsmodels MixedLM`, which is beyond the scope of this automated system's current capability. Because school identifiers are suppressed in the public-use file, the reconstruction cannot be verified against true school membership. Claims of generalisation to unseen schools are therefore provisional on the reconstruction approximating true school membership, which we cannot confirm.

A second limitation concerns survey weights. HSLS:09 uses a complex stratified multi-stage probability sampling design with analysis weights (`W1STUDENT`, `W2W1STU`, `W4W1W2W3STU`). The machine learning models were trained and evaluated without survey weights because scikit-learn's standard estimators do not support complex survey variance estimation that accounts for stratification and primary sampling unit clustering. Some models accept a `sample_weight` parameter, but using weights without proper variance estimation produces correctly weighted point estimates with incorrect standard errors, which would mislead readers about precision. The reported metrics therefore reflect unweighted sample performance and may not generalise exactly to the national population of ninth graders. Future work should use survey-aware machine learning packages, such as

weighted bootstrap procedures or the `survey` package in R, to produce properly weighted estimates.

Third, the analytic sample is conditioned on postsecondary entry (`X4EVRATNDCLG` = "Yes"). This is a deliberate design feature that makes the research question well-posed, but it means the findings do not generalise to all high school students, including those who never enrolled in postsecondary education. The selection process is non-random, and any inference about sector sorting applies only to the subpopulation of students who entered postsecondary education.

Fourth, missing data were assumed to be missing at random (MAR). The outcome variable `X4EVR2YPUB` has structural missingness (26.72% in the full dataset), and the analytic sample represents students with a valid outcome record. This is a population restriction, not random dropout. If missingness on the outcome is related to unobserved factors that also predict sector choice, the estimates could be biased. Sensitivity analyses using pattern-mixture models or multiple imputation under MNAR assumptions would strengthen confidence in the findings.

Fifth, the sex subgroup analysis could not be computed because `X1SEX` was not available in the protected attributes file for the test set. The research question promised fairness analysis across sex, race/ethnicity, and SES, but the sex-based fairness claim is only partially addressed. Future runs must ensure that `X1SEX` is carried into the protected attributes file so that sex subgroup performance can be computed. Relatedly, the race/ethnicity subgroup analysis included one cell with only 12 students (American Indian/Alaska Native, non-Hispanic), and the Native Hawaiian/Pacific Islander group (n = 9) was excluded entirely. The 11.2 percentage-point race gap in AUC is driven partly by these small-n cells and should not be over-interpreted.

Sixth, `X1SES` is used both as a predictor and as a subgroup attribute. Subgroup analyses for SES therefore require careful interpretation because SES is also in the model; the subgroup AUC differences reflect both the model's differential accuracy and the fact

that SES itself is a strong predictor of the outcome.

Finally, the design is correlational. No causal claims about the effect of any predictor on sector choice are warranted. The associations reported here are predictive, not explanatory, and should be interpreted accordingly.

References

Alsayed, K. A. (2026). Why aggregate accuracy is inadequate for evaluating fairness in law enforcement facial recognition systems [arXiv preprint].

Broden, J., Simson, J., & Kern, C. (2026). Conformal prediction and the many-worlds of fairness. *Conference on Fairness, Accountability and Transparency*. https://doi.org/10.1145/3805689.3812391

Brown, K. E., Yan, C., Li, Z., Zhang, X., Collins, B. X., Chen, Y., Clayton, E., Kantarcioglu, M., Vorobeychik, Y., & Malin, B. A. (2025). Large language models are less effective at clinical prediction tasks than locally trained machine learning models. *J. Am. Medical Informatics Assoc.* https://doi.org/10.1093/jamia/ocaf038

Carriero, A., Luijken, K., de Hond, A. D., Moons, K., van Calster, B., & van Smeden, M. (2024). The harms of class imbalance corrections for machine learning based prediction models: A simulation study. *Statistics in Medicine*. https://doi.org/10.1002/sim.10320

Davis, S. E., Dorn, C., Park, D., & Matheny, M. E. (2025). Emerging algorithmic bias: Fairness drift as the next dimension of model maintenance and sustainability. *J. Am. Medical Informatics Assoc.* https://doi.org/10.1093/jamia/ocaf039

Gao, R., & Ni, Q. (2025). Fairedunet: A novel adversarial network for fairer educational dropout prediction. *Scientific Reports*. https://doi.org/10.1038/s41598-025-13632-w

Glandorf, D., Lee, H. R., Orona, G., Pumptow, M., Yu, R., & Fischer, C. (2024). Temporal and between-group variability in college dropout prediction. *International Conference on Learning Analytics and Knowledge*. https://doi.org/10.1145/3636555.3636906

Guanin-Fajardo, J., Guaña-Moya, J., & Casillas, J. (2024). Predicting academic success of college students using machine learning techniques. *International Conference on Data Technologies and Applications*. https://doi.org/10.3390/data9040060

Guazzo, A., Atzeni, M., Idi, E., Trescato, I., Tavazzi, E., Longato, E., Manera, U., Chiò, A., Gromicho, M., Alves, I., de Carvalho, M., Vettoretti, M., & Camillo, B. D. (2024). Predicting clinical events characterizing the progression of amyotrophic lateral sclerosis via machine learning approaches using routine visits data: A feasibility study. *BMC Medical Informatics and Decision Making*. https://doi.org/10.1186/s12911-024-02719-5

Guevara-Reyes, R., Ortiz-Garcés, I., Andrade, R. O., Cox-Riquetti, F., & Villegas-Ch., W. (2025). Machine learning models for academic performance prediction: Interpretability and application in educational decision-making. *Frontiers in Education*. https://doi.org/10.3389/feduc.2025.1632315

He, S., Yousefpoori-Naeim, M., Cui, Y., & Cutumisu, M. (2025). Predicting college enrollment for low-socioeconomic-status students using machine learning approaches. *Big Data and Cognitive Computing*. https://doi.org/10.3390/bdcc9040099

Khilnaney, D., Maikner, M., Snyder, R. W., Chew, A., & Bailey, J. M. (2025). Beyond the money: Mentoring activities in support of two-year college stem majors. *Community College Journal of Research and Practice*. https://doi.org/10.1080/10668926.2025.2559647

Lünich, M., & Keller, B. (2024). Explainable artificial intelligence for academic performance prediction. an experimental study on the impact of accuracy and simplicity of decision trees on causability and fairness perceptions. *Conference on Fairness, Accountability and Transparency*. https://doi.org/10.1145/3630106.3658953

Mahawar, K., & Rattan, P. (2024). Empowering education: Harnessing ensemble machine learning approach and aco-dt classifier for early student academic performance prediction. *Education and Information Technologies : Official Journal of the IFIP technical committee on Education*. https://doi.org/10.1007/s10639-024-12976-6

Means, D. R. (2024). At the crossroads: Postsecondary education access opportunities and constraints for rural black students. *Review of higher education (Print)*. https://doi.org/10.1353/rhe.2025.a946305

Opoku, R. A., Pei, B., & Xing, W. (2025). Unveiling accuracy-fairness trade-offs: Investigating machine learning models in student performance prediction. *Journal of Learning Analytics*. https://doi.org/10.18608/jla.2025.8543

Plasman, J., & Myles, D. (2024). Stem pathway and college progression: The link between engineering cte and postsecondary outcomes for students with learning disabilities. *The Exceptional Child*. https://doi.org/10.1177/00144029241247034

Raftopoulos, G., Davrazos, G., & Kotsiantis, S. (2024). Fair and transparent student admission prediction using machine learning models. *Algorithms*. https://doi.org/10.3390/a17120572

Raftopoulos, G., Davrazos, G., & Kotsiantis, S. (2025). Evaluating fairness strategies in educational data mining: A comparative study of bias mitigation techniques. *Electronics*. https://doi.org/10.3390/electronics14091856

Schudde, L., Callahan, R. M., & Kwon, Y. (2025). Language and postsecondary trajectories: How "ever-english learner" status predicts college student pathways and outcomes. *Review of higher education (Print)*. https://doi.org/10.1353/rhe.2025.a954242

Sharma, M., & Mansotra, V. (2026). A dual attention-driven neural ensemble framework for student outcome prediction. *Educational Sciences International Conference*. https://doi.org/10.1109/ESIC68176.2026.11495867

Tang, Z., Jain, A., & Colina, F. E. (2024). A comparative study of machine learning techniques for college student success prediction. *Journal of Higher Education Theory and Practice*. https://doi.org/10.33423/jhetp.v24i1.6764

Tiruneh, S. A., Rolnik, D. L., Teede, H., & Enticott, J. (2025). Temporal validation of machine learning models for pre-eclampsia prediction using routinely collected

maternal characteristics: A validation study. *Comput. Biol. Medicine*. https://doi.org/10.1016/j.compbiomed.2025.110183

Wang, H., Zhang, M., Mai, L., Li, X., Bellou, A., & Wu, L. (2025). An effective multi-step feature selection framework for clinical outcome prediction using electronic medical records. *BMC Medical Informatics and Decision Making*. https://doi.org/10.1186/s12911-025-02922-y

Xia, H., & Chen, K. (2026). Bridging algorithmic prediction and teacher judgment: An explainable ai framework with institutional fairness calibration for early warning systems. *Frontiers in Education*. https://doi.org/10.3389/feduc.2026.1881571

Yang, C., Fridgeirsson, E. A., Kors, J., Reps, J., & Rijnbeek, P. (2024). Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *Journal of Big Data*. https://doi.org/10.1186/s40537-023-00857-7

Yao, C., Cortez, C., & Yu, R. (2025). Towards fair and privacy-aware transfer learning for educational predictive modeling: A case study on retention prediction in community colleges. *International Conference on Learning Analytics and Knowledge*. https://doi.org/10.1145/3706468.3706567

Zambrano, A. F., & Baker, R. S. (2024). Long-term prediction from topic-level knowledge and engagement in mathematics learning. *International Conference on Learning Analytics and Knowledge*. https://doi.org/10.1145/3636555.3636851

Table 1

Model comparison on the school-aware test partition ($n = 2{,}578$).

Model	AUC	Acc.	Prec.	Recall	F1
XGBoost	0.704	0.675	0.648	0.641	0.643
StackingEnsemble	0.705	0.671	0.642	0.631	0.634
RandomForest	0.699	0.661	0.630	0.619	0.622
LogisticRegression	0.699	0.667	0.635	0.621	0.624
ElasticNet	0.692	0.644	0.636	0.646	0.634

Table 2

Grouped SHAP feature importance (mean absolute SHAP, aggregated across dummy columns).

Parent variable	Mean |SHAP|	Columns
X3TGPAACAD	0.517	1
X3THISCI	0.212	5
X1SES	0.186	1
X1REGION	0.180	3
X1TXMTSCOR	0.130	1
X1CONTROL	0.122	1
X1LOCALE	0.098	3
X3TCREDMAT	0.059	1
X1RACE	0.056	8
X3TCREDTOT	0.043	1
X3TCREDENG	0.033	1
X1SEX	0.013	2

Table 3

Subgroup AUC on the test partition.

Group	AUC	*n*
Race/ethnicity		
Asian, non-Hispanic	0.778	226
Black/African-American, non-Hispanic	0.709	244
Hispanic, race specified	0.700	353
Hispanic, no race specified	0.700	28
White, non-Hispanic	0.688	1,374
More than one race, non-Hispanic	0.669	227
Amer. Indian/Alaska Native, non-Hispanic	0.781	12
Socioeconomic status (quintiles)		
Highest quintile	0.731	474
Fourth quintile	0.674	474
Middle quintile	0.680	473
Second quintile	0.680	473
Lowest quintile	0.627	475

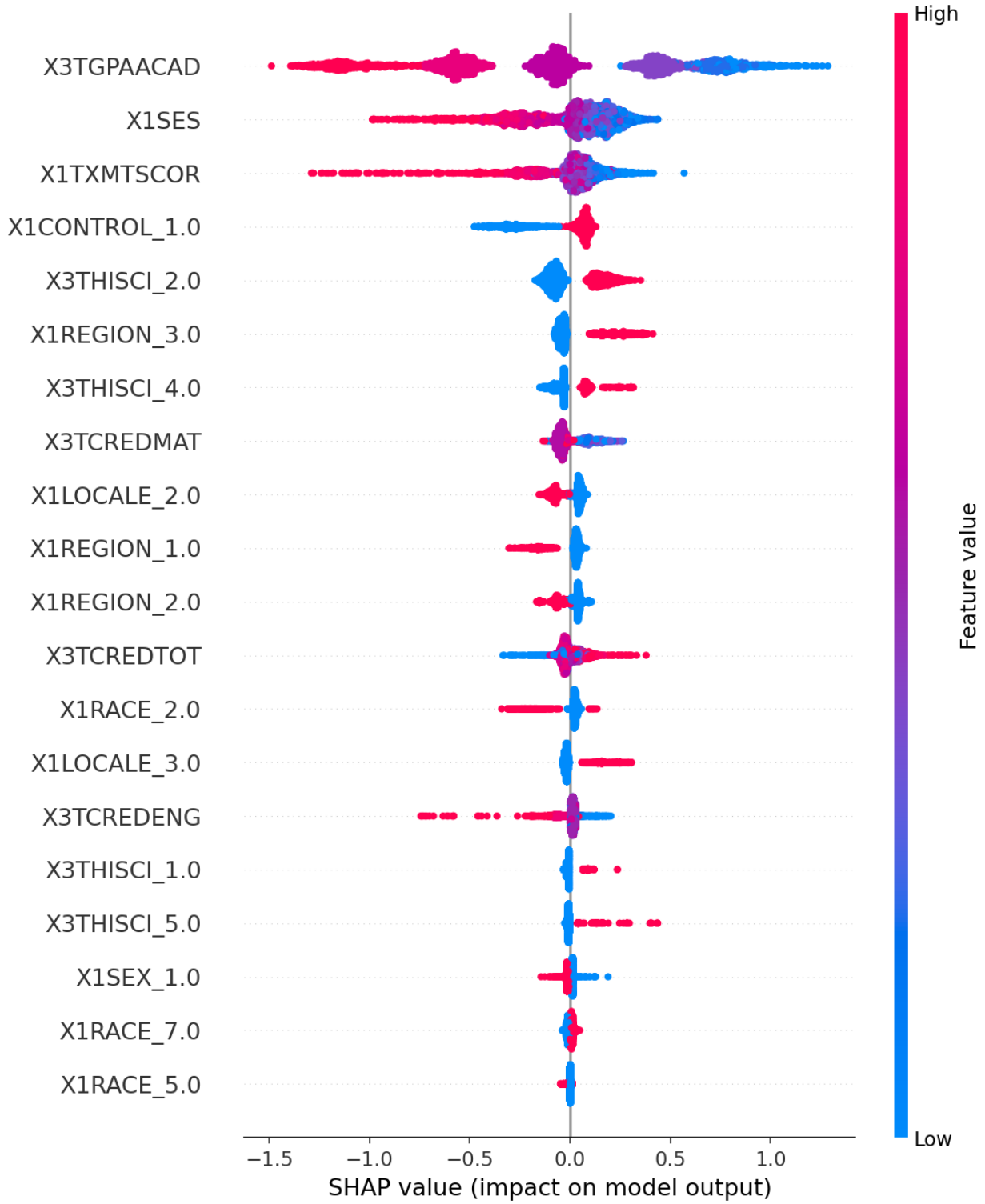

Figure 1

SHAP summary plot for the XGBoost model. Each point is one student; color indicates feature value.

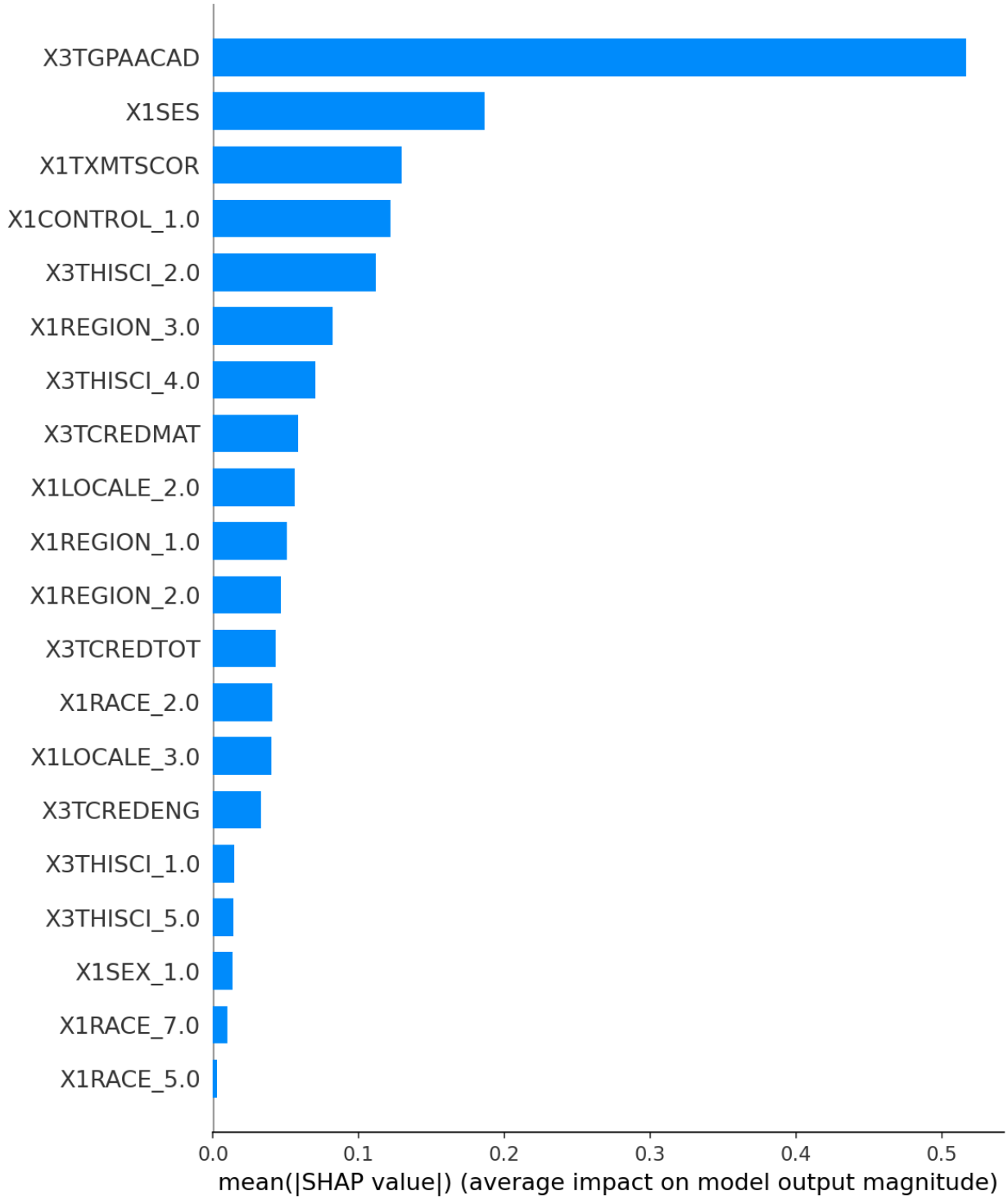

Figure 2

Mean absolute SHAP values for the top features.

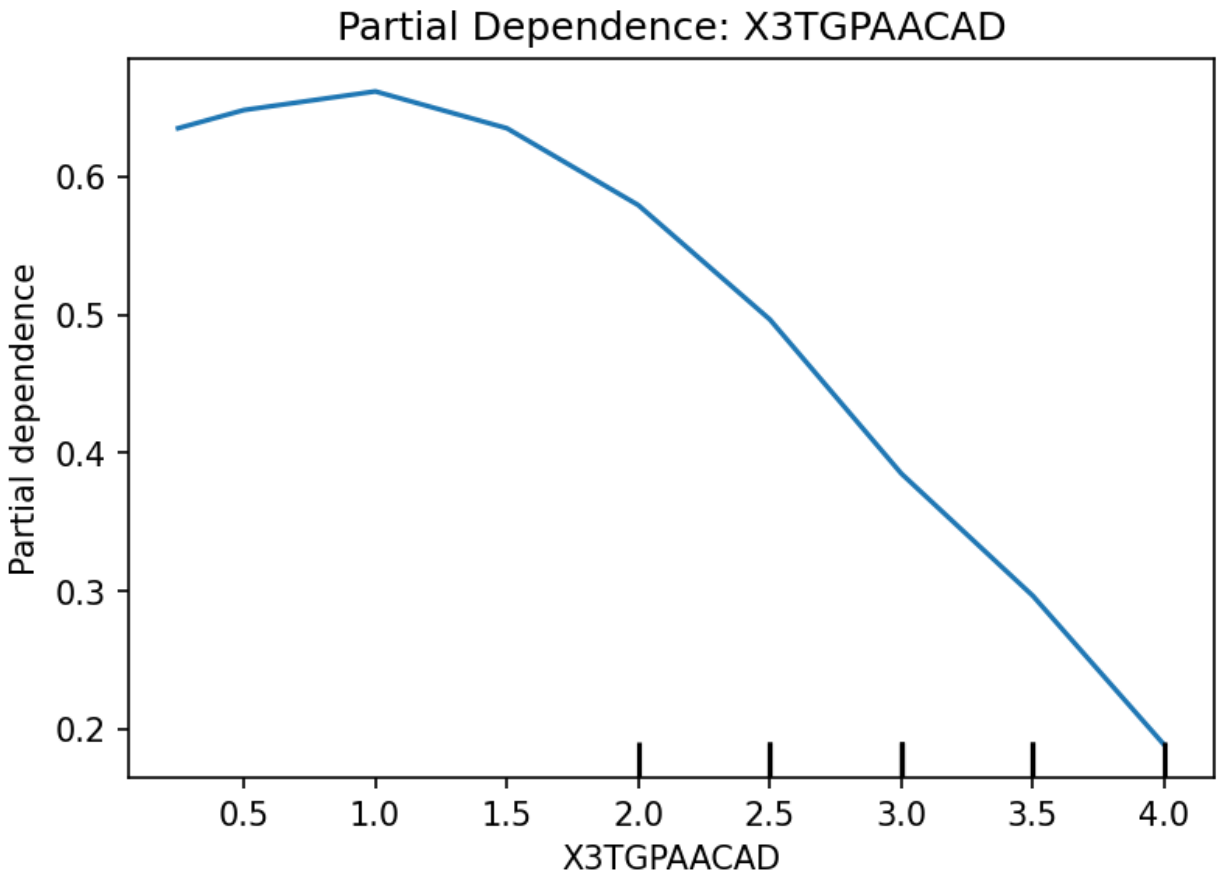

Figure 3

Partial dependence of predicted two-year public enrollment probability on academic GPA.

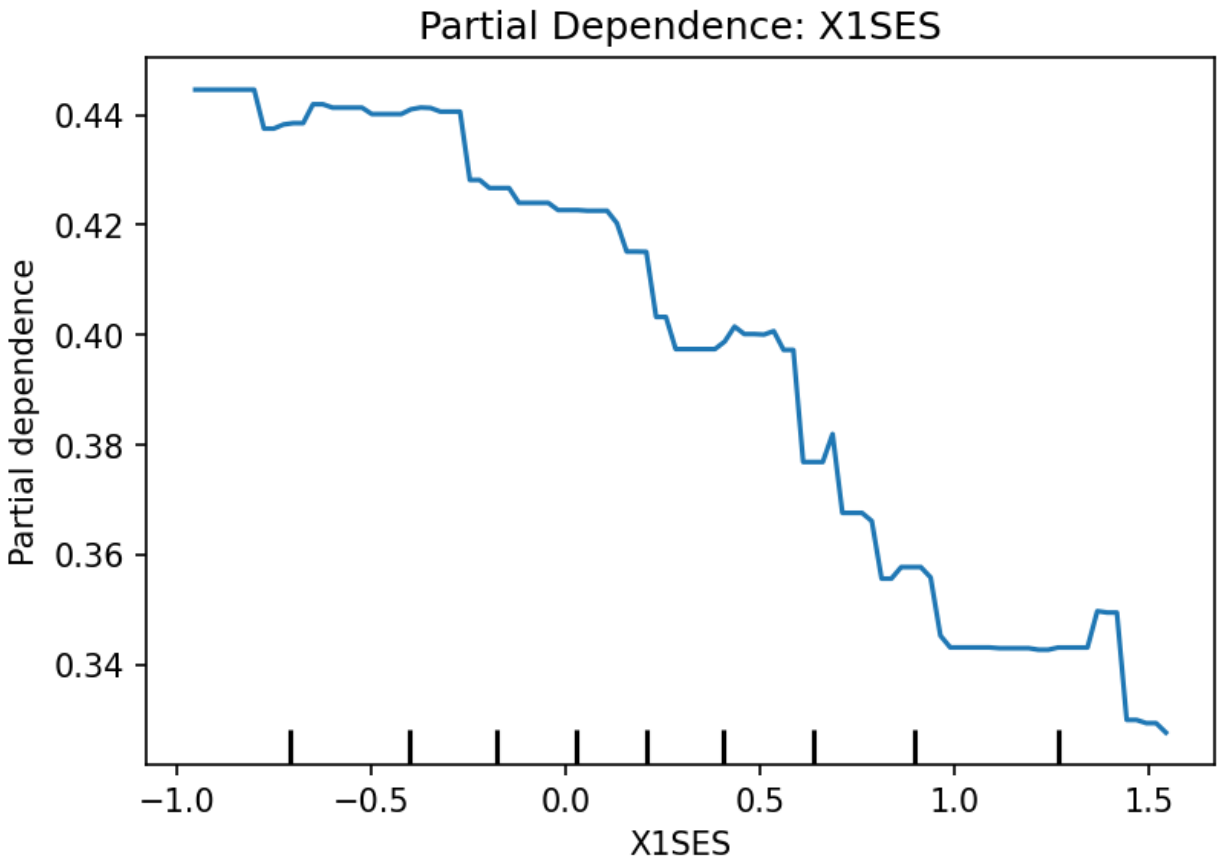

Figure 4

Partial dependence of predicted two-year public enrollment probability on socioeconomic status.

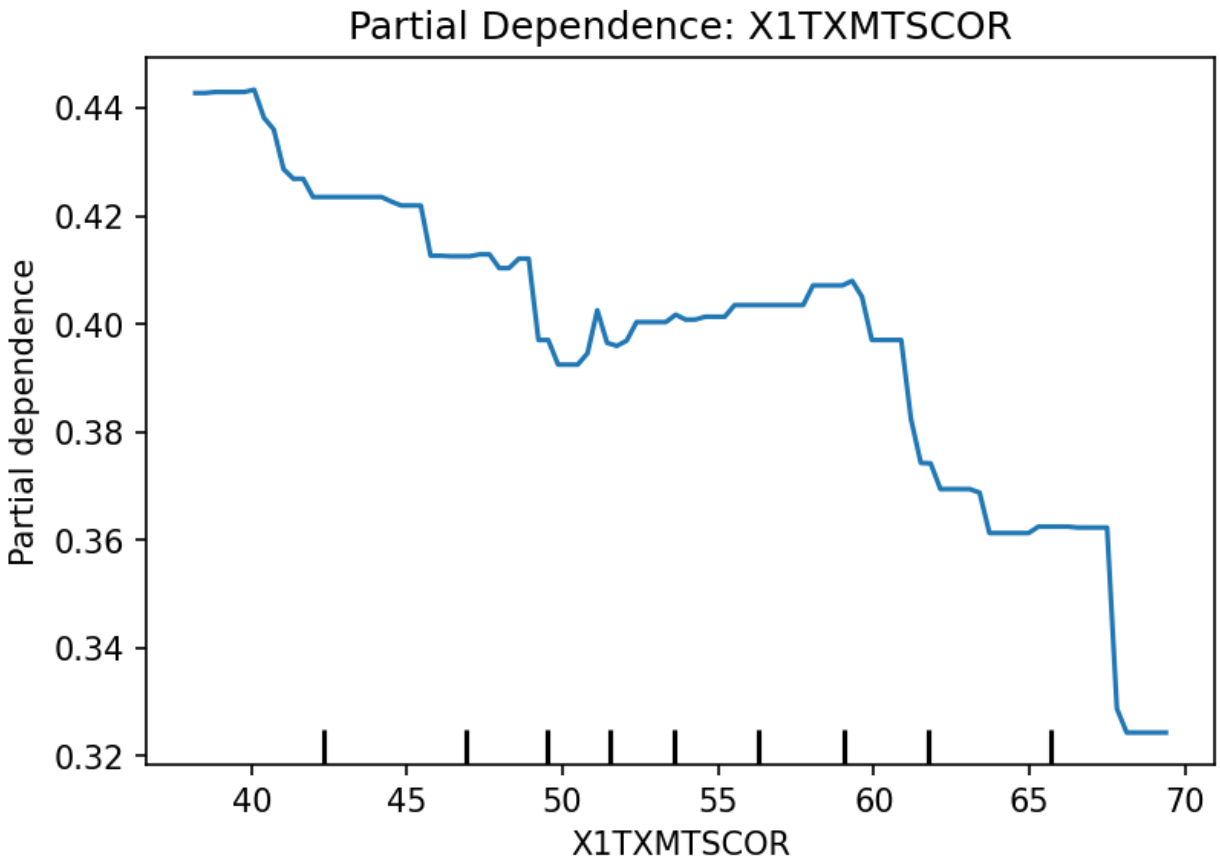

Figure 5

Partial dependence of predicted two-year public enrollment probability on ninth-grade mathematics achievement.

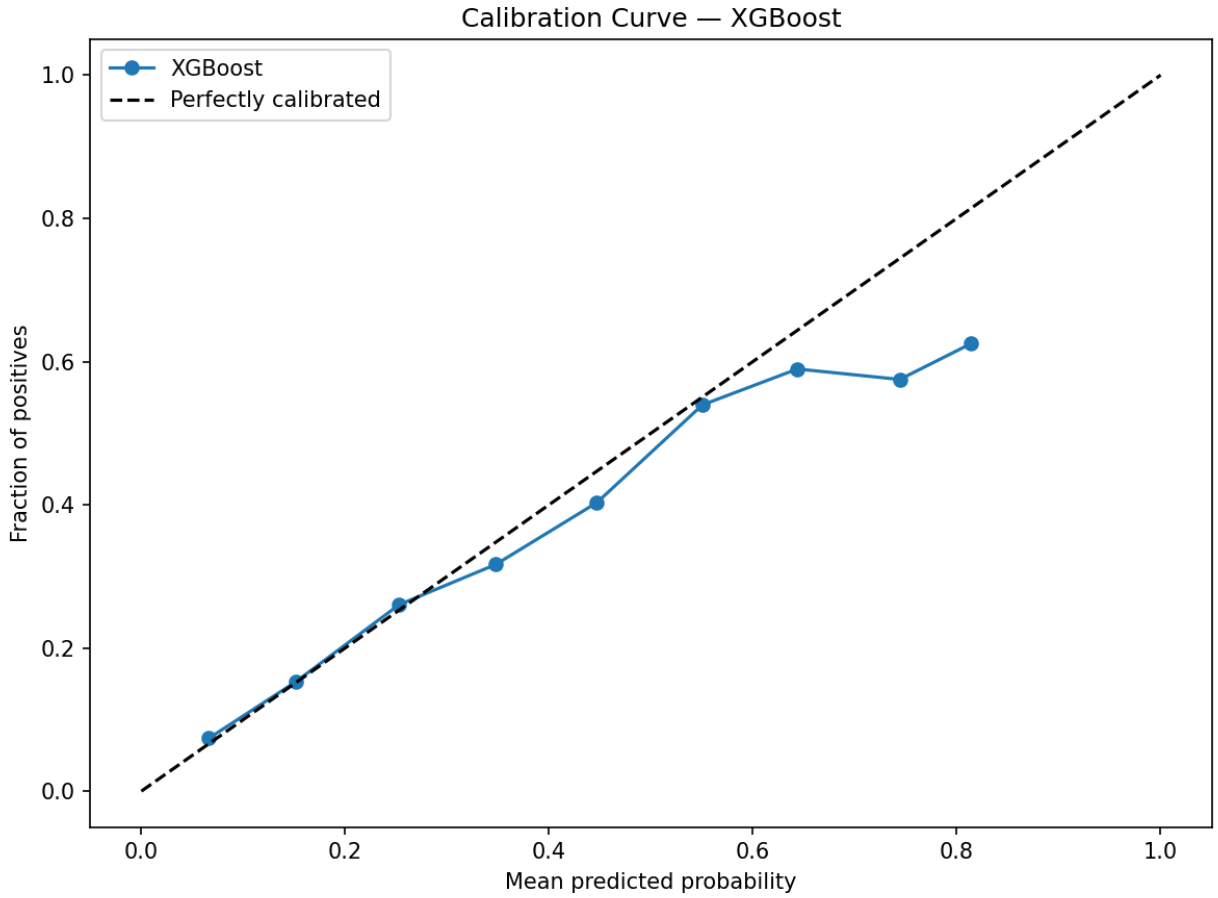

Figure 6

Calibration curve for the XGBoost model on the test partition.