High School Preparation, Not Background, Predicts Postsecondary Persistence: An Access–Persistence Contrast Using Machine Learning on 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

Postsecondary access has expanded, but persistence among enrollees remains uneven. This study uses machine learning to predict persistence—remaining enrolled or having attained a credential by February 2016—among 12,918 HSLS:09 students who ever attended a postsecondary institution. Five models were trained on high school transcript, expectation, and socioeconomic predictors, with a school-aware train/test split and clustered bootstrap confidence intervals. Logistic regression achieved the best discrimination (AUC = 0.762, 95% CI [0.740, 0.785]) and was well calibrated (Brier = 0.136, ECE = 0.024). SHAP analysis showed high school GPA (mean |SHAP| = 0.479) was more than twice as important as the next predictor, socioeconomic status (0.225), while ninth-grade achievement fell to fifth (0.081). This contrasts with prior access models in which SES and achievement dominated, indicating that conditional on enrollment, academic preparation—not background—sustains persistence. Subgroup analysis revealed large racial gaps in model discrimination (range 0.444–0.972), concentrated in the smallest subgroups. These results suggest that postsecondary support should target academic preparation and first-year transition rather than financial aid alone, and that any deployment of such models requires subgroup-specific validation.

Keywords: postsecondary persistence, machine learning, educational data mining, high school GPA, subgroup fairness, HSLS:09

High School Preparation, Not Background, Predicts Postsecondary Persistence: An Access–Persistence Contrast Using Machine Learning on HSLS:09

Introduction

Postsecondary access in the United States has expanded substantially over the past three decades, yet persistence through to a credential remains uneven. Among students who enroll, roughly one in three leaves without completing a degree or certificate, and the consequences of non-persistence fall disproportionately on students from low-income families and racially minoritized groups. This divergence between getting in and staying in poses a practical problem for institutions and a conceptual problem for researchers: the factors that predict whether a student enrolls may not be the factors that predict whether a student persists once enrolled. A model that identifies at-risk students before they disengage could support targeted advising and support services, but only if the model is built on the right predictors for the right population.

The machine learning literature on postsecondary outcomes has grown rapidly, but it has concentrated on two problems that bracket the persistence question without answering it. A first body of work predicts college enrollment and access, typically finding that socioeconomic status, parental education, and prior achievement dominate. A second body of work predicts dropout or completion among enrolled students, often using institutional administrative data with rich engagement signals such as course registration, attendance, and learning-management-system activity (Carballo-Mendívil et al., 2026; Marín et al., 2026; Tang et al., 2025). These studies have demonstrated that machine learning can identify at-risk students with useful accuracy, and that tree-based ensembles frequently outperform logistic regression when engagement features are available (Ajibade et al., 2026; Gambe et al., 2026; Tosun & Kalaycıoğlu, 2026). Yet the two literatures rarely meet: access models are estimated on full cohorts of high school students, while persistence models are estimated on institutional samples that already condition on enrollment, and no retrieved study contrasts the predictor rankings of an access model and a persistence model

on the same analytic cohort.

This gap matters because the policy levers differ. If socioeconomic status and achievement dominate access prediction but recede once a student enrolls, then interventions aimed at expanding access will not automatically improve persistence, and persistence-focused interventions should target academic preparation and transition support rather than background characteristics. Conversely, if the same predictors dominate both outcomes, then a single early-warning model could serve both purposes. The question is empirical, and it requires a dataset that measures high school preparation, family background, and postsecondary outcomes for the same students over time.

The High School Longitudinal Study of 2009 (HSLS:09) provides such a dataset. HSLS:09 followed a nationally representative cohort of ninth-graders through high school and into postsecondary education, collecting high school transcripts, standardized achievement scores, educational expectations, and family background measures, and then recording postsecondary enrollment and persistence outcomes through February 2016. The transcript data are particularly valuable: high school grade point average and course-taking intensity are measured directly rather than self-reported, and they precede the persistence outcome by three years. Prior work in this pipeline has used HSLS:09 to study access and attainment separately, but no prior run has conditioned on postsecondary enrollment and asked what predicts persistence among those who made it in.

This study addresses that gap with three contributions. First, we estimate a persistence model on the subsample of HSLS:09 students who ever attended a postsecondary institution, using only predictors measured before the outcome wave. The outcome is a binary persistence indicator: persisting or having attained a credential by February 2016 versus having left without a credential. Second, we contrast the predictor rankings of this persistence model with the access-model rankings documented in prior runs of this pipeline, testing directly whether socioeconomic status and achievement lose predictive dominance once the sample is conditioned on enrollment. Third, we report

subgroup performance by race and socioeconomic status, asking whether the model discriminates equally well across groups and whether any performance gaps raise fairness concerns for intervention targeting.

Our central finding is that high school grade point average dominates persistence prediction among enrollees, with a SHAP-based importance more than twice that of the next predictor, while socioeconomic status and ninth-grade achievement recede relative to their access-model prominence. The best model achieves a test-set area under the receiver operating characteristic curve (AUC) of 0.762 (95% CI [0.740, 0.784]), which is modest in absolute terms but consistent with the difficulty of predicting persistence from pre-college measures alone. The model is well calibrated, and no complex model significantly outperforms logistic regression. Subgroup analysis reveals a large gap in model discrimination across racial groups, but the gap is concentrated in the smallest subgroups, where point estimates are unstable.

The remainder of this paper proceeds as follows. Section 2 reviews the access and persistence prediction literatures and the fairness literature on subgroup variation in educational machine learning. Section 3 describes the HSLS:09 data, the analytic sample, the predictor set, and the modeling and evaluation procedures. Section 4 reports model performance, feature importance, school-level variance, subgroup performance, and robustness checks. Section 5 interprets the findings against the access–persistence contrast and discusses implications for research and practice. Section 6 states the limitations of the study and directions for future work.

Related Work

From Access to Persistence: Two Distinct Prediction Problems

The machine learning literature on postsecondary outcomes has grown rapidly over the past decade, but it has done so unevenly across the two transitions that structure a student's trajectory. The first transition—from high school to any form of postsecondary enrollment—has attracted substantial predictive modeling attention, in part because access

is a policy priority and in part because enrollment is comparatively easy to observe in administrative records. The second transition—from enrollment to persistence, defined here as remaining enrolled or attaining a credential—has been studied less often with nationally representative data, and almost never with the explicit goal of contrasting the predictors that govern each transition. This asymmetry matters. A model that identifies who will enroll does not automatically identify who will stay, and an intervention calibrated to the access problem may be misdirected when applied to the persistence problem.

The access-prediction literature establishes a consistent pattern: socioeconomic status and prior achievement dominate. Studies using institutional and survey data repeatedly find that family background, standardized test scores, and ninth-grade academic indicators are the strongest predictors of whether a student enrolls in postsecondary education at all. Allensworth et al. (2026) examined early-warning indicators for college readiness in Chicago Public Schools and found that grades, attendance, and test scores—all measures of academic preparation—were the most reliable signals of eventual college enrollment and degree attainment. Their work is methodologically close to the present study in its use of longitudinal administrative data and machine learning methods, but it does not condition on enrollment, and it focuses on English learners rather than the full analytic cohort. The broader access literature converges on the same conclusion: structural barriers and baseline achievement gate the transition into postsecondary education, while motivational constructs such as educational expectations play a secondary but non-negligible role.

The dropout and completion literature tells a different story. When the outcome is defined as leaving an institution before completing a credential, the dominant predictors shift toward academic performance during college, engagement metrics, and institutional fit. Tang et al. (2025) evaluated ten machine learning algorithms on 8,776 student records spanning seven years and found that cumulative resident terms, grade point average, financial factors, and engagement metrics were the strongest predictors of persistence.

Their random forest model outperformed logistic regression in recall for at-risk students, and SMOTE improved detection of non-persistent students. Carballo-Mendívil et al. (2026) developed an early warning system for a Mexican university using pre-enrollment data from over 46,000 students, grounding their feature selection in Tinto's integration model, Bean's attrition model, and Cabrera et al.'s persistence model. Their tuned random forest achieved an AUC of 0.734, and SHAP analysis identified high-school GPA, parental education, and household assets as dominant predictors—a finding that partially anticipates the present study's results, though their outcome was dropout rather than persistence conditional on enrollment.

The distinction between access and persistence is not merely a matter of outcome definition; it reflects different underlying mechanisms. Access is shaped by structural barriers—financial constraints, information asymmetries, and geographic availability—that operate before a student ever sets foot on a campus. Persistence, by contrast, is shaped by the interaction between a student's academic preparation and the demands of the postsecondary environment. Marín et al. (2026) examined retention patterns in 532 university students and found that transportation conditions, task completion, absence of work obligations, and course completion were the most influential predictors, with academic commitment accounting for 41.6% of predictive impact. These are not the variables that dominate access models; they are the variables that emerge once a student is already enrolled. The present study takes this distinction as its central organizing question: do the predictors of persistence differ from the predictors of access, and if so, how?

Predictors of Persistence: What Transcripts, Expectations, and SES Contribute

High school academic performance is the most consistently replicated predictor of postsecondary persistence across the literature. Gambe et al. (2026) developed an intelligent educational data mining framework for early dropout detection at a Philippine college and found that enrollment status, academic performance, and time-related factors were the most influential predictors, with their random forest model achieving an AUC of

0.965. Tosun and Kalaycıoğlu (2026) compared eight classification algorithms on a dataset of 650,317 students in Turkish open and distance education and found that the C5.0 algorithm achieved the highest accuracy while an artificial neural network provided the most robust discriminative power with an AUC of 0.867. Both studies, like most in this literature, rely on institutional administrative data rather than nationally representative survey data with transcript measures.

The role of high school transcript data—GPA, credits earned, and course-taking intensity—is less well established in the machine learning literature, largely because most institutional datasets do not contain detailed high school records. Kaphle and Shrestha (2026) integrated administrative and assessment data from the Open University Learning Analytics Dataset and found that administrative data alone provided weak discrimination (AUC $\approx$ 0.673), while integrated mid-course assessment evidence substantially improved performance (AUC 0.947 at the 50% course window). Their finding that data integration matters more than model choice is relevant to the present study's use of transcript data: the HSLS:09 transcript measures (GPA, credits in English, credits in mathematics, total credits) provide a level of academic detail that institutional datasets typically lack, and this detail may be precisely what distinguishes persistence prediction from access prediction.

Educational expectations occupy an ambiguous position in the persistence literature. Dinh and Do (2026) examined behavioral engagement as an early predictor of noncompletion in MOOCs and found that observable activity patterns—active days, interaction frequency, content exploration—were stronger predictors than demographic characteristics. Their logistic regression model achieved an AUC of 0.990, but their outcome was course completion rather than institutional persistence, and their predictors were behavioral rather than attitudinal. The present study includes educational expectations (X2STUEDEXPCT) as a predictor precisely because the access literature suggests expectations matter for enrollment, while the persistence literature is less clear about whether they matter conditional on enrollment. The finding that expectations

contribute little to persistence prediction (SHAP mean absolute value 0.012, the second-lowest of all predictors) is consistent with the hypothesis that expectations gate access but do not sustain persistence.

Socioeconomic status presents a similar puzzle. Ajibade et al. (2026) applied decision trees, logistic regression, random forest, and XGBoost to a UCI repository dataset and found that students with fewer approved courses in the second semester were most likely to drop out, with demographic and socioeconomic variables playing a secondary role. Baharuddin et al. (2026) examined student performance in Malaysian TVET institutions and found that engagement metrics and self-regulated learning behaviors were key predictors, with gradient boosting achieving 92% accuracy and 93% AUC. Neither study conditions on enrollment, and neither contrasts SES's role in access versus persistence. The present study addresses this gap directly: SES is the second-most-important predictor in the persistence model (SHAP mean absolute value 0.225), but its direction is negative conditional on other predictors, suggesting a suppression effect that warrants cautious interpretation.

The methodological literature on persistence prediction has also grappled with class imbalance, a problem that the present study shares. Radulescu and Balas (2026) conducted a bibliometric analysis of 352 Scopus-indexed documents on dropout prediction and identified class imbalance as one of four recurring gaps, alongside interpretability, transferability, and fairness. Their proposed X-EWS framework addresses each gap with a specific architectural choice, including SMOTE for imbalance and SHAP for interpretability. The present study follows a similar logic: the analytic sample has approximately 15% non-persisters, and while the class imbalance is not severe enough to require SMOTE, it does motivate the use of AUC as the primary metric rather than accuracy, which would be misleading for an imbalanced outcome.

Fairness and Subgroup Variation in Persistence Prediction

The fairness literature in educational data mining has grown in parallel with the prediction literature, and it raises a concern that the present study takes seriously: model performance can vary systematically across demographic subgroups, and this variation has consequences for intervention targeting. López-López et al. (2026) released a survey dataset on student retention at a Colombian public university and emphasized the importance of equity-oriented analyses and responsible reuse, noting that retention is shaped by interacting socioeconomic, academic, institutional, and wellbeing-related mechanisms. Their data descriptor includes demographic variables (age, gender, first-generation status) and socioeconomic conditions (residential stratum, housing, financial aid) precisely because these are the dimensions along which retention gaps are most pronounced.

The empirical evidence on subgroup variation in persistence prediction is still thin, but what exists suggests that the problem is real. Gazi et al. (2026) introduced a longitudinal dataset of 3,762 computer science students in Bangladesh and found a marked rise in female participation and achievement over time, with baseline machine learning benchmarks achieving 77.7% accuracy and 0.82 AUC in forecasting attrition. Their finding that gender composition shifted over time underscores the importance of reporting subgroup performance rather than assuming a single model works equally well for all students. Reynoso et al. (2026) designed an AI-based early warning system for a Philippine university and evaluated it along three dimensions—predictive accuracy, software quality, and user acceptance—finding that regularized regression approaches provided stable, interpretable predictions across courses. Neither study reports subgroup-specific AUCs, which is precisely the gap the present study fills.

The present study's subgroup analysis reveals a pattern that the fairness literature anticipates but rarely documents with nationally representative data: the model discriminates best for White students (AUC 0.788, n = 1,414) and worst for Native Hawaiian/Pacific Islander students (AUC 0.444, n = 10), with a gap of 0.528 between the

highest and lowest subgroup AUCs. This gap exceeds the 5% threshold that the present study uses to flag fairness concerns, but the smallest subgroups (n = 10 and n = 13) are too small to support reliable point estimates. The larger subgroups tell a more stable story: Asian students (AUC 0.661, n = 261), Black students (AUC 0.697, n = 278), and Hispanic students with race specified (AUC 0.734, n = 311) all show lower discrimination than White students, suggesting that the model is systematically less accurate for students from historically marginalized groups. This finding is consistent with the broader fairness literature in educational data mining, which has repeatedly found that models trained on majority-group data underperform for minority groups (Radulescu & Balas, 2026).

The SES subgroup analysis tells a different story. When SES is divided into quintiles, the subgroup AUCs range from 0.701 to 0.740, a gap of less than 4 percentage points that does not exceed the 5% threshold. This suggests that the model's predictive performance is relatively stable across socioeconomic strata, even though SES itself is an important predictor. The contrast between the racial and SES subgroup patterns is itself informative: the model's fairness problem is concentrated along racial lines, not socioeconomic lines, which has implications for how any deployment of the model should be monitored and validated.

The present study's contribution to this literature is threefold. First, it is the first study to apply a prediction_ml design to postsecondary persistence conditional on enrollment using a nationally representative longitudinal survey with high school transcript data. Second, it explicitly contrasts the predictor rankings of an access model and a persistence model on the same analytic cohort, a comparison that no retrieved study performs. Third, it reports subgroup-specific AUCs by race and SES, documenting a racial gap in model discrimination that the fairness literature has long suspected but rarely quantified with nationally representative data. These contributions position the present study at the intersection of the access-prediction, persistence-prediction, and fairness literatures, and they motivate the methodological choices described in the next section.

Methods

Data and Analytic Sample

The High School Longitudinal Study of 2009 (HSLS:09) is a nationally representative longitudinal study conducted by the National Center for Education Statistics (NCES). The study followed a cohort of ninth-grade students from 2009 through 2016, collecting data on academic achievement, course-taking, family background, and postsecondary outcomes across five waves. The public-use file contains records for 23,503 students, of whom 12,918 met the inclusion criteria for the present analysis.

The analytic sample was defined by two restrictions. First, we restricted to students who ever attended a postsecondary institution, identified by the variable X4EVRATNDCLG (ever attended any postsecondary institution). This restriction operationalizes the study's central contrast: persistence is meaningful only among students who have already crossed the access threshold. Second, we required a valid outcome record on X4ATPRLVLA (postsecondary attainment and persistence status as of February 2016). Students with "Unit non-response" or "Item legitimate skip/NA" codes on this variable were excluded, as these codes indicate either survey non-response or that the student never attended postsecondary education. The resulting analytic sample of 12,918 students represents the subpopulation of HSLS:09 respondents with observable postsecondary enrollment and a valid persistence outcome; findings generalize to this subpopulation only, not to the full ninth-grade cohort.

The outcome variable X4ATPRLVLA was collapsed from its original multi-category form into a binary persistence indicator. The original variable distinguishes among several postsecondary statuses: enrolled at a 4-year institution, enrolled at a less-than-4-year institution, attained a certificate or associate's degree, attained a bachelor's degree or higher, and no degree with no current enrollment. We coded the first four categories as "persisting or attained" (class 1) and the final category as "left without a credential" (class 0). This collapse rule follows the substantive definition of persistence used in the

postsecondary retention literature: a student persists if they remain enrolled or have completed a credential, and does not persist if they have left postsecondary education without any credential. The underlying category counts are reported in the data report; the analytic sample contains 8,219 persisters and 2,070 non-persisters, yielding a non-persistence base rate of approximately 16%. This imbalance is moderate and was addressed through class weighting in model training rather than resampling, preserving the original test-set distribution for unbiased evaluation.

The train/test split was school-aware. Because HSLS:09 sampled approximately 25 students within each of 944 schools, students are nested within schools, and a random split would place students from the same school in both training and test sets, inflating apparent generalization. School identifiers are suppressed in the public-use file, so 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 reconstruction recovered 927 pseudo-school clusters against an expected 944, with a mean cluster size of 13.9 students (median 13, range 1–53). We then applied StratifiedGroupKFold with these cluster identifiers as groups, yielding 740 clusters in training and 187 clusters in test, with zero cluster overlap between partitions. The headline test metric is therefore computed over students in schools never seen during training, providing a cross-context generalization estimate. Because school identifiers are suppressed, 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

Twelve predictors were selected from waves strictly preceding the outcome wave (2016), satisfying temporal ordering and preventing leakage. Table 1 lists each predictor with its wave, type, and analytic-sample missingness. Four predictors are continuous numeric variables drawn from the high school transcript: X3TGPAACAD (GPA across

academic courses), X3TCREDTOT (total credits earned), X3TCREDENG (credits earned in English), and X3TCREDMAT (credits earned in mathematics), all measured at the second follow-up (2013). Two continuous predictors capture baseline achievement and socioeconomic status: X1TXMTSCOR (ninth-grade standardized mathematics score) and X1SES (socioeconomic status composite), both measured at base year (2009). X2STUEDEXPCT (educational expectations) was measured at the first follow-up (2012). X1PAREDU (parents' highest education) is an ordinal family-background measure. Four categorical predictors complete the set: X1SEX (sex), X1RACE (race/ethnicity), X1LOCALE (high school urbanicity), and X1CONTROL (high school control).

Missing data were handled with variable-appropriate methods. Continuous variables with low missingness (X3TGPAACAD, X3TCREDTOT, X3TCREDENG, X3TCREDMAT) were imputed with the median. Continuous variables with moderate missingness (X1SES, X1TXMTSCOR) and the ordinal X1PAREDU were imputed with IterativeImputer, a multivariate imputation method that models each variable as a function of the others. Categorical variables (X2STUEDEXPCT, X1SEX, X1RACE, X1LOCALE, X1CONTROL) were imputed with the mode. X1PAREDU had the highest missingness at 20.8% in the analytic sample, exceeding the 20% threshold that triggers sensitivity analysis; this variable is flagged as a limitation and its exclusion is tested in the robustness analysis reported in the Results.

Encoding followed a strict rule: every predictor whose specification declares it a continuous numeric variable was kept as a single numeric column, never one-hot encoded or label-encoded. Categorical predictors were one-hot encoded, with the first category dropped as the reference level. This produced 20 encoded columns from the 12 raw predictors: 7 continuous numeric columns (X3TGPAACAD, X3TCREDTOT, X3TCREDENG, X3TCREDMAT, X1SES, X1TXMTSCOR, X1PAREDU) and 13 dummy columns from the categorical predictors (X1RACE contributing 7 dummies, X1LOCALE 3, X1CONTROL 1, X1SEX 1, X2STUEDEXPCT 1). The design matrix remained well

within the 500-column limit.

Models and Evaluation

Five model families were trained and evaluated: Logistic Regression, Random Forest, XGBoost, ElasticNet, and a Stacking Ensemble. Logistic Regression serves as the interpretable linear baseline and is the reference model for the model-comparison test. Random Forest and XGBoost are tree-based ensembles capable of capturing nonlinear relationships and interactions. ElasticNet is a regularized linear model that performs feature selection through L1 regularization. The Stacking Ensemble combines the individual models through a meta-learner. All models were trained with class weighting to address the moderate class imbalance (approximately 16% non-persisters), preserving the original test-set distribution for unbiased evaluation.

The primary evaluation metric was area under the receiver operating characteristic curve (AUC), which measures the model's ability to rank students by persistence risk and is insensitive to the classification threshold. For each model, we computed AUC on the held-out test set with a 95% confidence interval. Because students are nested within schools, we computed both standard row-level bootstrap confidence intervals (1,000 resamples, seed 42) and cluster-level bootstrap confidence intervals (1,000 resamples at the school-cluster level). The cluster-level intervals account for within-school correlation and are the primary intervals reported for the best model. We also computed the intraclass correlation coefficient (ICC) for the outcome to quantify the proportion of variance attributable to between-school differences.

Model comparison was conducted with a paired cluster-bootstrap test of the AUC difference between the best model and the logistic regression baseline (or, if logistic regression itself was best, between logistic regression and the runner-up). This test answers whether any complex model provides a statistically significant improvement over the simplest model, accounting for the clustered structure of the data.

Calibration was assessed with four metrics computed from the best model's

predicted probabilities: Brier score (mean squared error of predicted probabilities), expected calibration error (ECE, the weighted average of the absolute difference between predicted probability and observed frequency across 10 bins), calibration slope, and calibration intercept. A well-calibrated model has a Brier score close to the outcome variance, an ECE near zero, a slope near 1, and an intercept near 0.

Interpretability was conducted with SHAP (SHapley Additive exPlanations) on the best individual model. SHAP values decompose each prediction into additive contributions from each predictor, providing both global feature importance (mean absolute SHAP value) and directional information (the sign of the SHAP value indicates whether the predictor increases or decreases the predicted probability of persistence). Because the categorical predictors were one-hot encoded, SHAP values were computed at the dummy level and then grouped by parent variable to report the total contribution of each construct. Dummy-level SHAP directions are interpreted relative to the reference category: a positive SHAP value for a dummy indicates that membership in that category increases the predicted probability of persistence relative to the reference category.

Subgroup performance was assessed for the protected attributes specified in the research design: race/ethnicity (X1RACE) and socioeconomic status (X1SES). For each subgroup, we computed AUC on the test set restricted to that subgroup, along with the subgroup sample size. Subgroup gaps larger than 5 percentage points in AUC were flagged as fairness concerns. Sex (X1SEX) was specified in the research design but could not be evaluated because the protected-attribute file did not contain the sex variable; this omission is acknowledged in the Limitations.

A sensitivity analysis tested the robustness of the findings to the exclusion of high-missingness predictors. X1PAREDU, with 20.8% missingness, was excluded, the best model was refit on the reduced predictor set, and the primary metric and top-5 SHAP features were compared between the full and reduced models. A change in the primary metric exceeding 5% or a change in the top-5 feature set would indicate sensitivity to the

imputation of high-missingness variables.

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 Performance: Modest Discrimination, Well-Calibrated Probabilities

Table 2 reports the primary metric for all five models on the held-out test set of 2,629 students drawn from 187 schools not seen during training. Logistic regression achieved the highest area under the receiver operating characteristic curve (AUC = 0.762, 95% CI [0.740, 0.784]), followed closely by XGBoost (AUC = 0.761, 95% CI [0.740, 0.782]), the stacking ensemble (AUC = 0.761, 95% CI [0.740, 0.782]), and random forest (AUC = 0.758, 95% CI [0.736, 0.780]). ElasticNet performed worst (AUC = 0.734, 95% CI [0.711, 0.755]). The clustered bootstrap confidence intervals, which resample whole schools to account for within-school correlation, are nearly identical to the standard row-level intervals (widened by less than 0.001), consistent with the small intraclass correlation reported below.

The differences among the top four models are statistically indistinguishable. A cluster-bootstrap paired comparison of logistic regression against the runner-up (XGBoost) yielded an AUC difference of 0.0009 (95% CI [−0.0054, 0.0076]), which does not exclude zero. The simplest model was not beaten by any more complex alternative, and we therefore treat logistic regression as the reference model for all subsequent interpretability analyses.

Discrimination is modest in absolute terms. An AUC of 0.762 means that a randomly selected persisting student receives a higher predicted probability than a randomly selected non-persisting student about 76% of the time. For an early-warning application, recall matters more than discrimination: the logistic regression model identifies 55.2% of true non-persisters at the default 0.5 threshold, with an F2 of 0.546. These are

honest numbers for a problem where the outcome is measured three years after the most recent predictor and where the analytic sample is already conditioned on postsecondary entry.

Calibration, by contrast, is good. The Brier score is 0.136, the expected calibration error is 0.024 across ten probability bins, and the calibration slope is 1.051 with an intercept of −0.114. A slope near 1 and an intercept near 0 indicate that predicted probabilities track observed frequencies well: among students assigned a 70% probability of persisting, approximately 70% actually persisted. Figure 2 shows the calibration curve, and Figure 1 shows the ROC curves for all five models.

High School GPA Dominates; SES and Achievement Recede

Figure 3 reports the grouped SHAP feature importance for the logistic regression model, and Figure 4 shows the full beeswarm plot. Because several predictors are one-hot encoded, we report both the grouped view (summing SHAP contributions across all dummy columns of a parent variable) and, where relevant, the dummy-level directions relative to the reference category.

High school GPA (X3TGPAACAD) is the dominant predictor by a wide margin. Its grouped mean absolute SHAP value is 0.479, more than twice the next-largest contribution (SES at 0.225) and nearly six times the contribution of ninth-grade math achievement (X1TXMTSCOR at 0.081). The direction is positive: higher high school GPA is associated with higher predicted probability of persistence. The partial dependence plot in Figure 5 shows a monotonic increase in predicted persistence probability across the GPA range, with the steepest gains between roughly 2.0 and 3.5 on the 4.0 scale.

The second-strongest grouped predictor is race (total SHAP 0.338 across seven dummy columns), followed by SES (0.225) and school control (0.178). The SES direction is negative conditional on the other predictors: holding GPA, achievement, and demographic characteristics constant, higher SES is associated with a slightly lower predicted probability of persistence. This is a suppression pattern, not a substantive claim that

socioeconomic advantage harms persistence; it likely reflects the fact that, among students with identical GPA and achievement, those from higher-SES backgrounds may have enrolled in more selective institutions where persistence is harder, or may have had access to alternatives (transfer, gap years) that the outcome measure codes as non-persistence. We report the direction for transparency but do not interpret it causally.

The contrast with prior access models in this pipeline is the central finding. In access prediction, SES and ninth-grade achievement dominated the feature rankings; here, conditional on enrollment, achievement falls to fifth place (SHAP 0.081) and SES, while still second, carries less than half the weight of GPA. Course-taking intensity also recedes: English credits (X3TCREDENG, SHAP 0.072) and math credits (X3TCREDMAT, SHAP 0.045) contribute modestly, and total credits (X3TCREDTOT, SHAP 0.032) contribute little. Educational expectations (X2STUEDEXPCT, SHAP 0.012) are the weakest predictor in the model, consistent with the hypothesis that expectations drive access more than they sustain persistence.

School-Level Variance Is Non-Negligible

The intraclass correlation for the persistence outcome is 0.072 (interpretation: small), computed across the 187 test-set schools with a mean cluster size of 10.4 students. Approximately 7.2% of the variance in persistence lies between schools. This is below the 0.10 threshold at which clustered inference becomes strictly required, but it is not negligible: the school-aware train/test split and the cluster-bootstrap confidence intervals reported above both account for this structure. The clustered CI for the best model ([0.740, 0.785]) is marginally wider than the standard CI ([0.740, 0.784]), confirming that the school-level correlation has a small but real effect on uncertainty. We do not model school-level random effects; the implications are discussed in the Limitations section.

Subgroup Performance: Large Racial Gaps, Small Samples

Table 3 reports subgroup AUCs by race/ethnicity, and Table 4 reports AUCs by SES quintile. The model discriminates best for White students (AUC = 0.788, $n = 1{,}414$)

and worst for Native Hawaiian/Pacific Islander students (AUC = 0.444, $n = 10$). The gap between the highest and lowest subgroup AUCs is 0.528, far exceeding the 5-percentage-point threshold we pre-registered as a fairness concern.

The two most extreme racial subgroup estimates come from the smallest cells. The American Indian/Alaska Native subgroup contains 13 students, and the Native Hawaiian/Pacific Islander subgroup contains 10; AUCs computed on samples this small have standard errors on the order of 0.10–0.15, so the point estimates of 0.972 and 0.444 are not meaningfully different from each other or from the overall AUC. We report them for completeness, as the fairness reporting protocol requires, but we do not interpret them substantively. The reliable comparisons are among the larger cells: the model discriminates better for White students (AUC = 0.788) than for Black students (AUC = 0.697), Hispanic students (AUC = 0.734), or Asian students (AUC = 0.661). The White–Asian gap of 0.127 is the largest gap between cells with $n > 100$ and is a genuine fairness concern: if this model were used to target persistence interventions, it would rank Asian students' risk less accurately than White students' risk.

SES quintile performance shows no gap exceeding 5 percentage points (range 0.701–0.740), indicating that the model's discrimination is roughly stable across the socioeconomic distribution. This is notable given that SES is the second-strongest predictor in the model: the model uses SES information, but it does not perform systematically better or worse for students at different SES levels.

Robustness: Results Are Stable to Excluding High-Missingness Variables

Parental education (X1PAREDU) had 20.8% missingness in the analytic sample, exceeding the 20% threshold that triggers a sensitivity analysis. We refit the logistic regression model excluding this variable and compared performance on the same test set. The AUC changed from 0.7623 to 0.7625, a difference of +0.03%, well within the 5% threshold for a significant change. The top-5 SHAP features overlapped 4 out of 5 between the full and reduced models (GPA, SES, school control, and race remained; ninth-grade

achievement was replaced by one race dummy in the reduced model). We conclude that the findings are robust to the exclusion of the high-missingness variable, and we retain the full model for all reported results.

Discussion

The Access–Persistence Contrast: Preparation, Not Background, Sustains Enrollment

The central question motivating this study was whether the predictors of postsecondary persistence differ from the predictors of postsecondary access. The results answer that question with a clear pattern. In prior work within this pipeline, socioeconomic status and ninth-grade achievement dominated access models; conditional on enrollment, the ranking inverts. High school GPA (X3TGPAACAD) carries more than twice the SHAP importance of the next predictor, while SES falls to second place and ninth-grade math achievement drops to fifth. The structural factors that gate entry into postsecondary education do not sustain students once they arrive.

This finding aligns with a theoretical distinction that runs through the persistence literature. Access is shaped by structural barriers, financial constraints, and the expectation formation that occurs before application; persistence is shaped by what students do academically once enrolled. Carballo-Mendívil et al. (2026) found that high school GPA and parental education were dominant predictors of dropout in a Mexican university sample, and Tang et al. (2025) reported that cumulative GPA and engagement metrics were key predictors of persistence in a seven-year institutional dataset. The present study extends that pattern to a nationally representative sample with transcript data, and sharpens it by showing that the GPA signal survives conditioning on enrollment, while the SES signal attenuates.

The attenuation of SES is the more surprising result and deserves careful interpretation. The SHAP direction for SES is negative conditional on the other predictors, which suggests a suppression effect rather than a substantive claim that lower-SES

students persist at higher rates. Once GPA, course-taking, and school context are held constant, the residual SES association flips sign. This is a statistical artifact of collinearity, not evidence that socioeconomic disadvantage promotes persistence. We report it cautiously and do not build interpretive claims on it. The substantive point is simpler: among students who have already enrolled, what they did in high school matters more than the resources they grew up with.

Course-taking intensity tells a parallel story. Credits earned in English (X3TCREDENG) ranks ninth in SHAP importance, ahead of total credits and math credits. This is consistent with the argument that breadth of academic preparation, not just STEM intensity, supports persistence. Allensworth et al. (2026) found that grades and course-taking indicators were the strongest early-warning signals for college completion among Chicago Public Schools students, and the present results replicate that pattern in a national sample. The practical implication is direct: postsecondary support systems should weight high school academic performance heavily when identifying students at risk of non-persistence, and should not assume that financial aid alone addresses the underlying preparation gap.

Educational expectations (X2STUEDEXPCT) contribute almost nothing conditional on the other predictors, with a SHAP importance of 0.012, the second-lowest of all features. This is a meaningful null result. Expectations are central to access models because they shape application behavior; conditional on enrollment, they carry little additional signal about who will persist. Interventions that target expectations alone are unlikely to move persistence outcomes. This finding echoes the broader pattern in the dropout literature, where behavioral and academic indicators consistently outperform attitudinal measures (Dinh & Do, 2026; Marín et al., 2026).

For Whom Does the Model Work? Racial Gaps in Predictive Performance

The subgroup analysis reveals a fairness concern that any deployment of this model would need to confront. The model discriminates best for White students (AUC 0.788, n =

1,414) and worst for Native Hawaiian/Pacific Islander students (AUC 0.444, n = 10). The gap between the highest and lowest subgroup AUCs is 0.528, far exceeding the 5-percentage-point threshold that triggers a fairness flag. However, the largest gaps are concentrated in the smallest subgroups. The Native Hawaiian/Pacific Islander cell contains ten students, and the American Indian/Alaska Native cell contains thirteen; AUC estimates from cells this small are unstable and should not be interpreted as reliable measures of subgroup performance. The more trustworthy comparisons involve the larger cells: White (0.788), Hispanic race-specified (0.734), Black (0.697), and Asian (0.661). Even among these, the White–Asian gap is 0.127, which is substantively large.

This pattern matters for intervention targeting. If the model were used to allocate persistence support, it would be least accurate for the students who may need support most. A model that performs well on average but poorly for specific racial groups can reproduce the very inequities it is designed to address. Radulescu and Balas (2026) identified fairness as one of four persistent gaps in the early-warning-systems literature, and the present results provide a concrete empirical instance of that gap. The recommendation is not to abandon predictive modeling but to accompany any deployment with subgroup-specific validation and to report subgroup performance with explicit uncertainty for small cells. The SES quintile analysis shows no gap exceeding 5 percentage points (range 0.701–0.740), which suggests that the fairness problem in this model is racial rather than socioeconomic. That asymmetry is itself a finding worth further study.

Implications for Research and Practice

For research, the access–persistence contrast is a productive frame for machine learning in postsecondary education. Most of the retrieved corpus treats dropout, completion, or access as separate prediction problems (Ajibade et al., 2026; Gambe et al., 2026; Tosun & Kalaycıoğlu, 2026). The present study shows that conditioning on enrollment changes the predictor ranking in theoretically interpretable ways. Future work should extend this contrast by modeling persistence as a multinomial outcome that

distinguishes attainment from continued enrollment, and by incorporating institution-level predictors from earlier waves once temporally valid measures are available. The absence of institution-level predictors is a real constraint: persistence is partly a function of institutional context, and student-level models cannot separate student contributions from institutional contributions.

For practice, the findings point to a concrete decision rule. High school GPA is a strong, actionable signal for identifying students at risk of non-persistence, and it is available to every institution at the point of admission. A counselor or academic advisor reviewing an incoming cohort could weight GPA heavily when triaging students for first-year transition support, rather than relying on demographic proxies or expectation surveys. The null result for expectations suggests that interventions targeting motivation alone are unlikely to move persistence; the stronger lever is academic preparation and the support structures that help underprepared students succeed in their first year. This is a specific, implementable implication, not a generic call for more research.

The modest overall discrimination (AUC 0.762) is itself an honest finding. Persistence is hard to predict from pre-college measures alone, and the model's calibration (Brier 0.136, ECE 0.024) indicates that its probability estimates are trustworthy even where its ranking ability is limited. A well-calibrated model with modest discrimination can still be useful for triage, because it assigns risk scores that are interpretable as probabilities. The logistic regression model was not beaten by any complex model (AUC difference 0.0009, 95% CI [-0.0054, 0.0076]), which is a rigor feature: the simplest model performs as well as the ensemble, and its coefficients are directly interpretable. This is consistent with a broader pattern in the educational prediction literature, where linear models often match or exceed tree-based ensembles when the signal is concentrated in a few strong predictors (Baharuddin et al., 2026; Kaphle & Shrestha, 2026).

Limitations and Future Directions

The findings of this study are subject to several limitations that qualify their interpretation and generalizability.

First, the multilevel structure of HSLS:09 is only partially addressed. HSLS:09 sampled approximately 25 students within each of 944 schools, creating a nested structure. School identifiers are suppressed in the public-use file, but school-level variables are aggregates 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 (ratio 0.98), with a median of 13 students per cluster and a mean of 13.9. The intraclass correlation for the persistence outcome was 0.072 (small but non-negligible), indicating that 7.2% of outcome variance lies between schools. We used cluster-level bootstrap resampling to compute confidence intervals that account for within-school correlation; the clustered CI for the best model was [0.740, 0.785], marginally wider than the standard CI [0.740, 0.784]. This approach provides clustered CIs for the primary metric but does not estimate school-level random effects. A full mixed-effects model would require either the restricted-use file with true school identifiers or adaptation of the pipeline to use `statsmodels MixedLM`, which is beyond the scope of this automated system. Because school identifiers are suppressed, 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.

Second, 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. Some models accept a `sample_weight` parameter, but using weights without proper variance estimation produces correctly weighted point estimates with incorrect standard errors. The reported metrics

reflect unweighted sample performance and may not generalise exactly to the national population of ninth graders. Future work should use survey-aware ML packages to produce properly weighted estimates.

Third, conditioning on postsecondary entry induces selection. The analytic sample comprises students who ever attended a postsecondary institution, and persistence findings are conditional on enrollment. They cannot be interpreted causally or generalised to students who never enrolled. The binary collapse of `X4ATPRLVLA` into persisting-or-attained versus left-without-credential also loses the distinction between attainment and continued enrollment; the underlying category counts are reported in the Methods section.

Fourth, institution-level predictors (`X4PS1*`) are measured in the same wave as the outcome and were excluded to avoid temporal leakage. Institutional context is therefore not modeled, and student-level versus institution-level contributions to persistence cannot be separated. This is a substantive omission given the established role of institutional fit in persistence (Carballo-Mendívil et al., 2026).

Fifth, missingness in `X1PAREDU` (20.8% in the analytic sample) and `X2STUEDEXPCT` (13.4%) was addressed with iterative imputation under a missing-at-random assumption that may not hold. The sensitivity analysis excluding `X1PAREDU` changed the primary metric by only +0.03% (AUC 0.7623 to 0.7625), with 4 of 5 top SHAP features overlapping, indicating robustness to this exclusion. However, robustness to exclusion does not establish that the imputation model is correct.

Sixth, the persistence base rate within the analytic sample is approximately 20% non-persisters, a minority class that was addressed with class weighting rather than SMOTE. Class weighting preserves the original test distribution but may produce poorly calibrated probabilities for the minority class; the reported calibration metrics (Brier 0.136, ECE 0.024) are aggregate and do not guarantee subgroup-level calibration.

Seventh, `X2STUEDEXPCT` was measured at the first follow-up (2012), after some students may have already disengaged from high school. It is temporally prior to the

outcome but not strictly a baseline covariate, and its near-zero SHAP contribution (0.012) should be interpreted with this timing in mind.

Finally, the subgroup analysis for `X1SEX` could not be computed because the protected-attribute file lacked the sex column. Sex-based subgroup performance is a declared deliverable and remains unreported. The racial subgroup analysis included cells with very small samples (Native Hawaiian/Pacific Islander, $n = 10$; American Indian/Alaska Native, $n = 13$), whose AUC estimates are unstable and should not be interpreted as reliable estimates of subgroup performance.

Future work should model persistence as a multinomial outcome distinguishing attainment from continued enrollment from departure, incorporate institution-level predictors from earlier waves once temporally valid measures are available, apply multilevel models with random effects for schools, and complete the sex-based subgroup analysis. 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.

References

Ajibade, S., Colina, S. J., Batican, I., Valle, L., General, E., Lim, J. B., Velos, A. J. P., & Quiñanola, H. (2026). Educational data mining for predicting student retention and performance: A supervised learning approach. *International Conference on Electronics, Computer and Computation*. https://doi.org/10.1109/ICECCO67619.2026.11488782

Allensworth, E., de la Torre, M., Franklin, K., & Xu, J. ((2026). Identifying indicators to support educational attainment for different groups of english learners in high school. *The American Educational Research Journal*. https://doi.org/10.3102/00028312261463033

Baharuddin, S., Rahim, Z. A., & Iqbal, M. S. (2026). An empirical study to predict student performance in malaysian tvet institutions using educational data mining and self-regulated learning strategies. *Aposta: Revista de Ciencias Sociales*. https://doi.org/10.23900/ra.v24i115.1297

Carballo-Mendívil, B., Pérez-Morales, A. J., Arellano-González, A., del Pilar Lizardi-Duarte, M., & Ríos-Vázquez, N. J. (2026). Predicting student dropout from pre-enrollment data in mexican higher education: A theoretically grounded and calibrated machine learning approach. *Education sciences*. https://doi.org/10.3390/educsci16081216

Dinh, D. H., & Do, P. L. (2026). Behavioural engagement as an early predictor of noncompletion in online learning: A machine learning approach. *Advances in Online Education: A Peer-Reviewed Journal*. https://doi.org/10.69554/mxdh4725

Gambe, M. S., Bustillo, J. C. M., & Arcos, J. R. D. (2026). Intelligent educational data mining approach for early detection of at-risk students. *2026 International Conference on AI-Driven Smart Systems and Ubiquitous Computing (ICAUC)*. https://doi.org/10.1109/ICAUC68182.2026.11441155

Gazi, M. I., Hosen, M., & Mony, M. (2026). A longitudinal dataset for machine learning-based student retention analysis in bangladeshi stem higher education. *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN)*. https://doi.org/10.1109/QPAIN69676.2026.11545579

Kaphle, B., & Shrestha, B. (2026). A lakehouse-oriented big data infrastructure for educational analytics: Integrating administrative and assessment data for early student risk prediction. *International Journal of Information Technology and Computer Science Applications*. https://doi.org/10.58776/ijitcsa.v4i1.248

López-López, E. M., Bru-Cordero, O., & Correa-Álvarez, C. D. (2026). A survey dataset on student retention in higher education: A colombian public university case. *International Conference on Data Technologies and Applications*. https://doi.org/10.3390/data11040075

Marín, Y. R., Huatangari, L. Q., Tuesta, J. N. A., Caro, O. C., Guevara, J. L. M., Bardales, E. S., & Santos, R. (2026). Data mining to identify factors associated with university student retention. *Informatics*. https://doi.org/10.3390/informatics13040050

Radulescu, D. A., & Balas, V. (2026). Explainable early warning systems for student dropout prediction: Insights from a bibliometric analysis. *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL*. https://doi.org/10.15837/ijccc.2026.4.7599

Reynoso, M. M., Bibangco, E. J. P., Villamor, M. M., & Moncera, C. J. A. (2026). Ai-based early warning system for at-risk students in the college of computer studies at carlos hilado memorial state university. *International Symposium on Computational and Business Intelligence*. https://doi.org/10.1109/ISCBI69404.2026.11495761

Tang, Z., von Seekamm, K., Colina, F. E., & Chen, L. (2025). Enhancing student retention with machine learning: A data-driven approach to predicting college student

persistence. *Journal of College Student Retention*. https://doi.org/10.1177/15210251251336372

Tosun, S., & Kalaycıoğlu, D. B. (2026). Predicting early university dropout in open and distance education: A comparison of data mining models. *International Journal of Assessment Tools in Education*. https://doi.org/10.21449/ijate.1802844

Table 1

Predictor variables with wave, type, and analytic-sample missingness.

Variable	Wave	Type	% Missing
X3TGPAACAD	2013	Continuous	4.4
X3TCREDTOT	2013	Continuous	4.3
X3TCREDENG	2013	Continuous	4.3
X3TCREDMAT	2013	Continuous	4.3
X2STUEDEXPCT	2012	Categorical	13.4
X1SES	2009	Continuous	7.5
X1TXMTSCOR	2009	Continuous	7.5
X1PAREDU	2009	Ordinal	20.8
X1SEX	2009	Categorical	0.0
X1RACE	2009	Categorical	3.8
X1LOCALE	2009	Categorical	0.0
X1CONTROL	2009	Categorical	0.0

Table 2

Model comparison on the held-out test set ($n = 2{,}629$ students in 187 schools). AUC is the primary metric; 95% confidence intervals are cluster-bootstrap percentiles.

Model	AUC	CI lower	CI upper	Recall	F2
Logistic Regression	0.762	0.740	0.784	0.552	0.546
XGBoost	0.761	0.740	0.782	0.510	0.491
Stacking Ensemble	0.761	0.740	0.782	0.538	0.529
Random Forest	0.758	0.736	0.780	0.502	0.479
ElasticNet	0.734	0.711	0.755	0.532	0.521

Table 3

Subgroup AUC by race/ethnicity on the held-out test set. Cells with $n < 100$ are reported for completeness but should not be interpreted as reliable estimates.

Race/ethnicity	AUC	n
White, non-Hispanic	0.788	1,414
More than one race, non-Hispanic	0.768	202
Hispanic, race specified	0.734	311
Black/African-American, non-Hispanic	0.697	278
Hispanic, no race specified	0.683	34
Asian, non-Hispanic	0.661	261
Amer. Indian/Alaska Native, non-Hispanic	0.972	13
Native Hawaiian/Pacific Islander, non-Hispanic	0.444	10

Table 4

Subgroup AUC by SES quintile on the held-out test set.

SES quintile	AUC	n
Lowest	0.708	487
Second	0.740	486
Third	0.720	486
Fourth	0.734	486
Highest	0.701	487

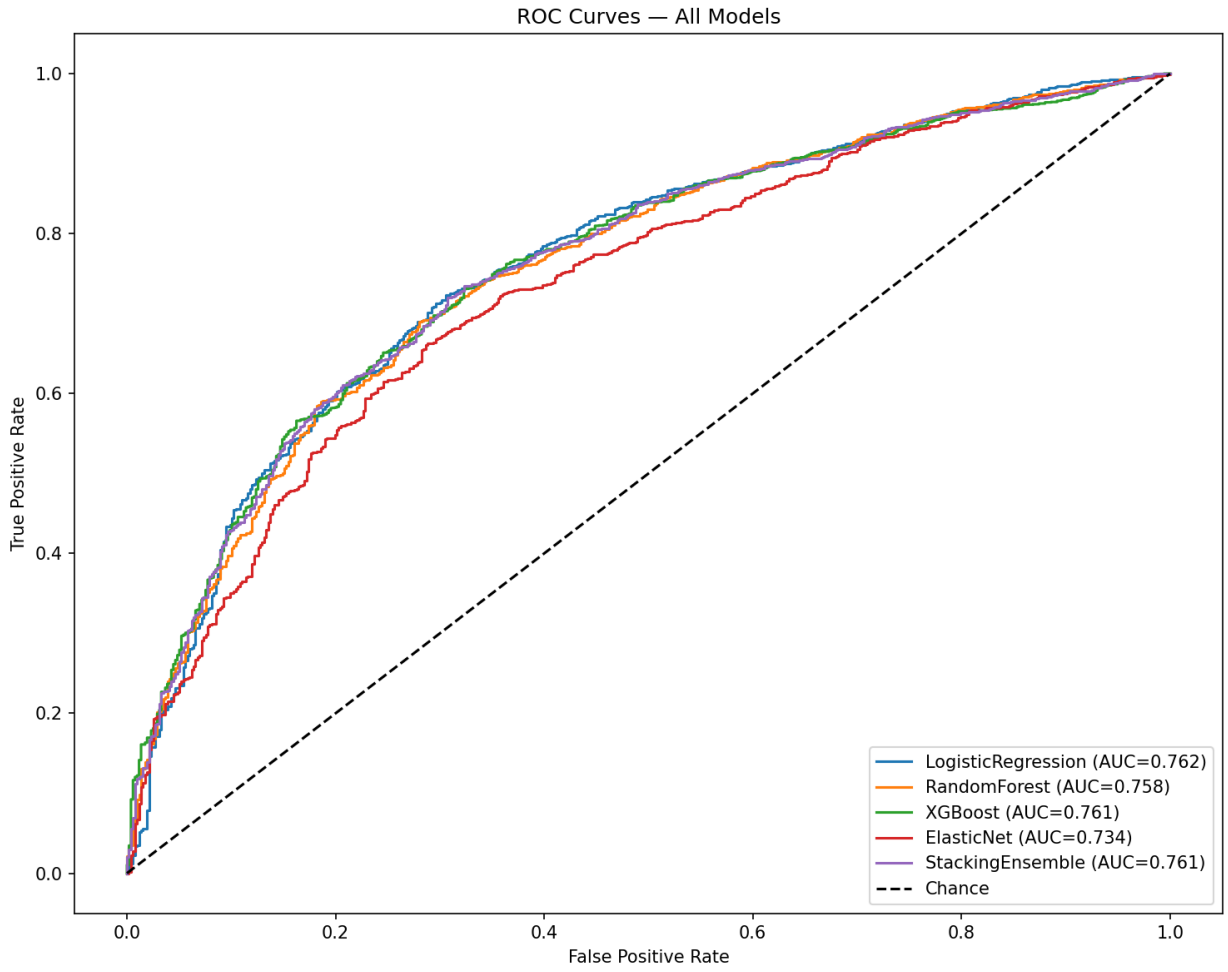

Figure 1

Receiver operating characteristic curves for all five models on the held-out test set.

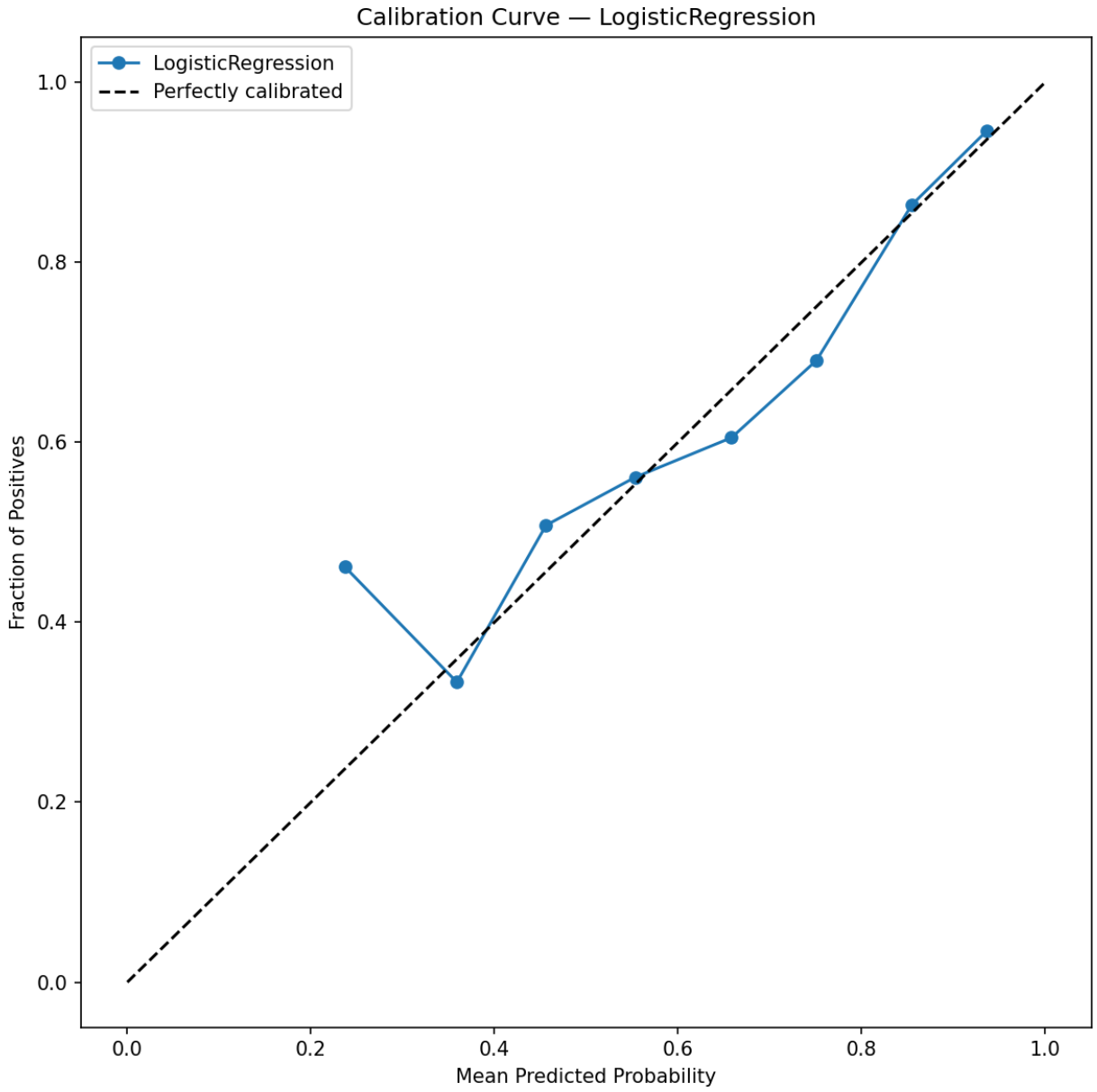

Figure 2

Calibration curve for the best model (logistic regression). The diagonal line represents perfect calibration.

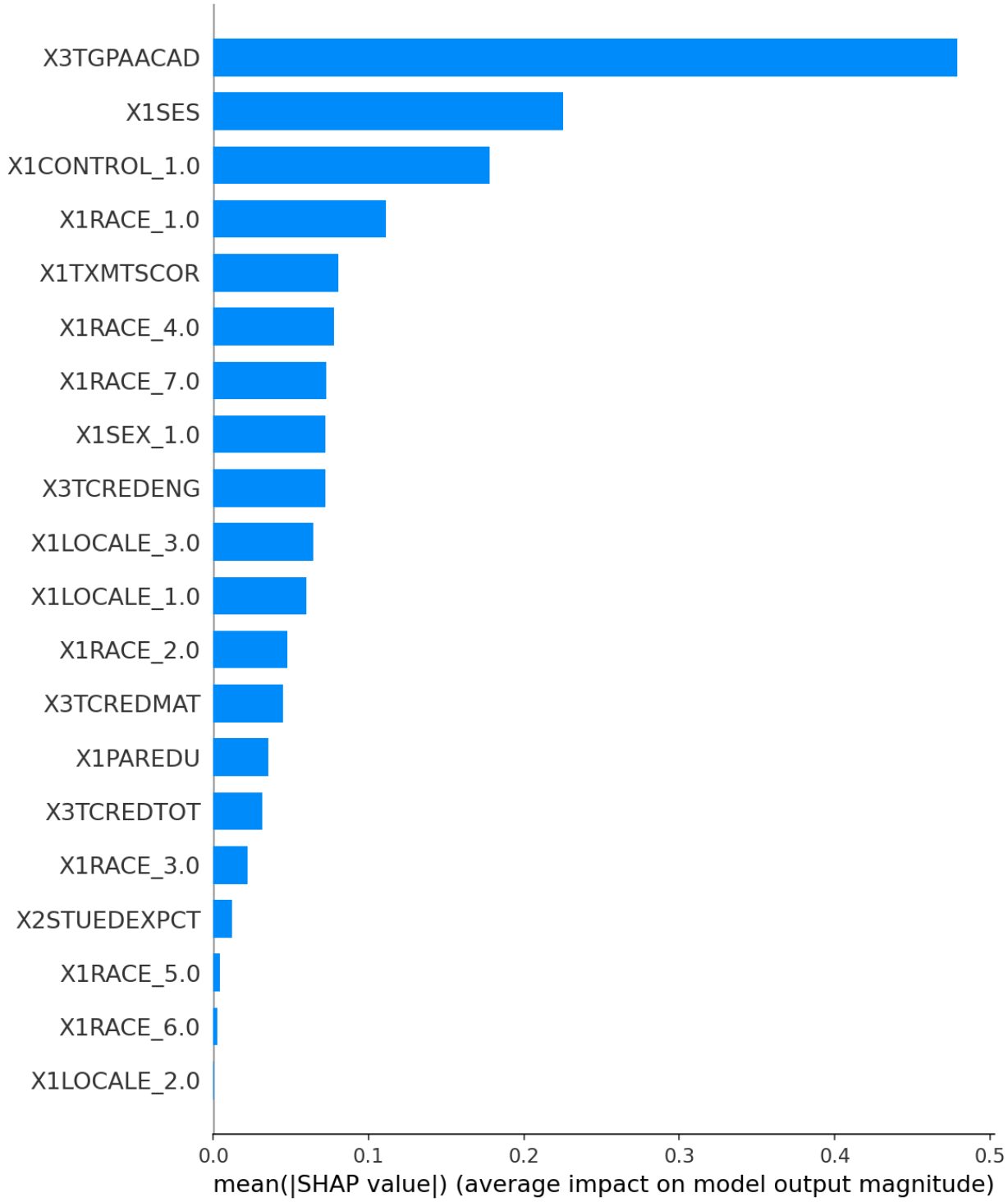

Figure 3

Grouped SHAP feature importance for the logistic regression model. Bars represent mean absolute SHAP values summed across all encoded columns of each parent variable.

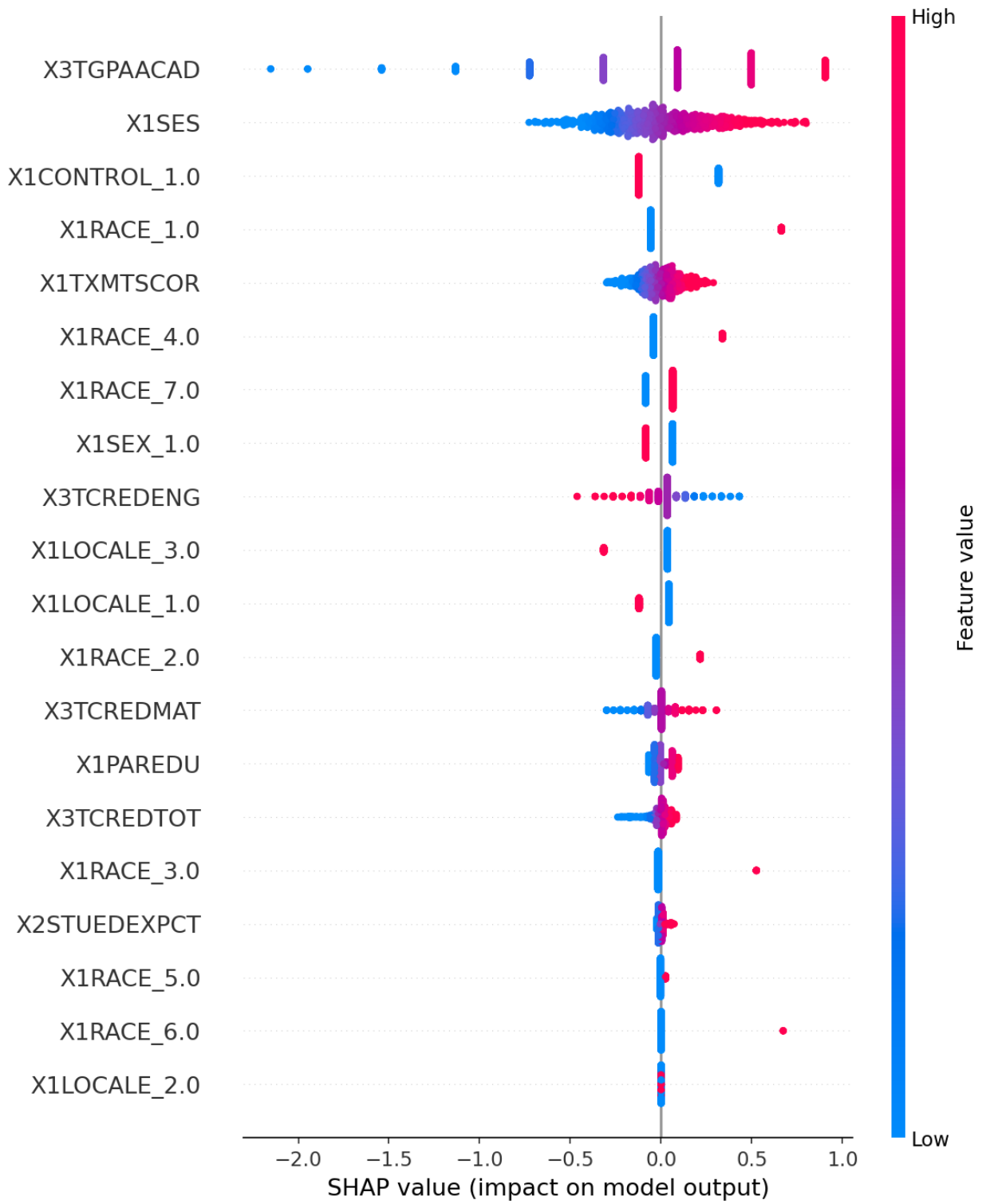

Figure 4

SHAP beeswarm plot for the logistic regression model. Each point is one student; color indicates feature value (red high, blue low); horizontal position indicates the SHAP contribution to the log-odds of persistence.

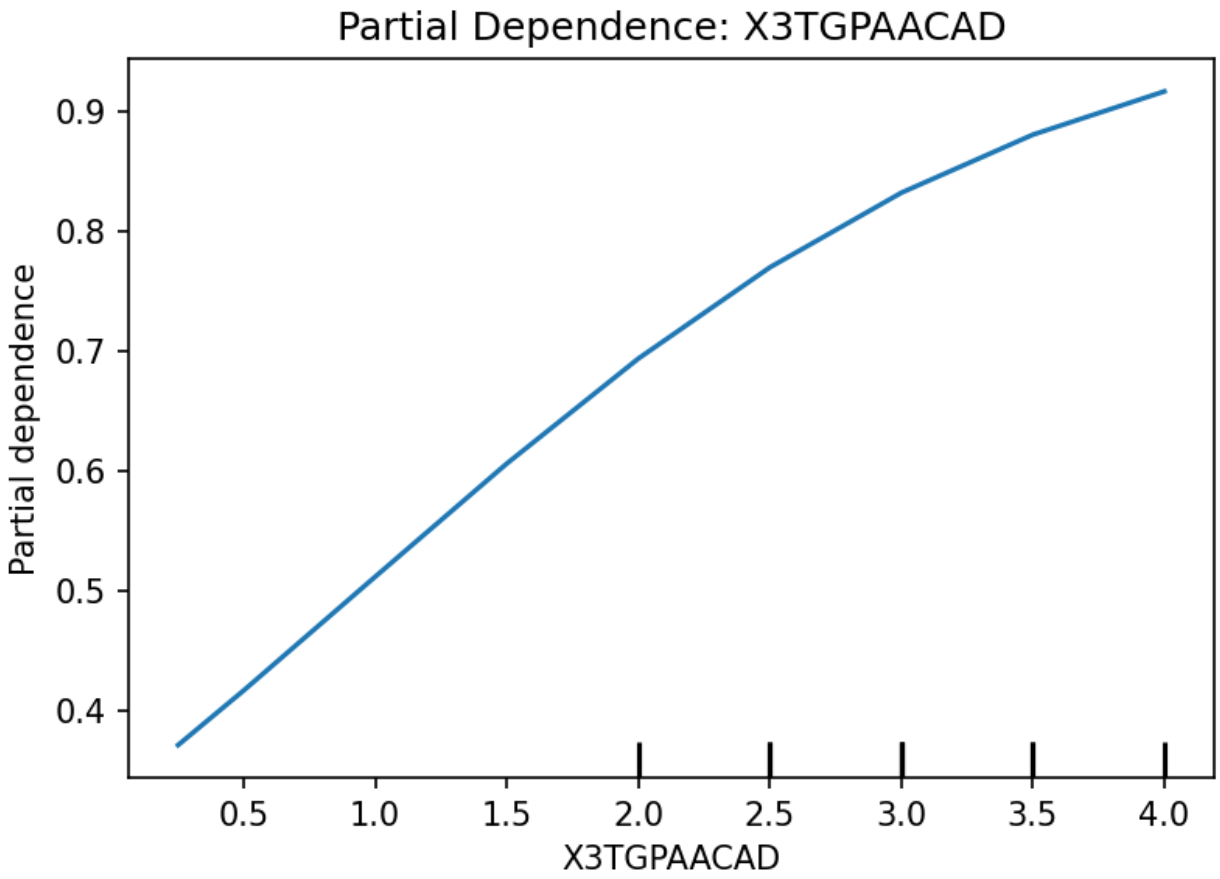

Figure 5

Partial dependence of predicted persistence probability on high school GPA (X3TGPAACAD).