Do Ninth-Grade Math and Science Identity Predict STEM Degree Non-Completion? Evidence from HSLS:09 and Machine Learning

EDM-ARS
edmars.ai
New York, NY, USA

ABSTRACT

Despite growing emphasis on diversifying the STEM workforce, educational data mining research on STEM outcomes has relied predominantly on short-term enrollment proxies or predictors from later high school waves, leaving the critical early intervention window underutilized. This study asks whether 9th-grade math and science identity, self-efficacy, school belonging, and engagement predict STEM degree non-completion beyond what math achievement and socioeconomic status explain, using a nationally representative cohort of U.S. students tracked from the High School Longitudinal Study of 2009 (HSLS:09) through postsecondary enrollment. Five machine learning models were trained on 11,560 students (9,161 training, 2,399 test) with school-aware cross-validation, and SHAP interpretability analysis was applied to the best individual model. The stacking ensemble achieved the highest AUC of 0.744 (95% CI [0.720, 0.767]), though all models performed within a narrow 5-point AUC band, indicating robust and consistent signal. Math and science identity ranked among the top predictors, rivaling math achievement itself: science identity (rank 3) and math identity (rank 4) together accounted for substantial predictive weight. Subgroup analysis revealed a 5.1-point AUC gap by gender and an 8.3-point gap by race, signaling that early warning systems built on these predictors may perform unevenly across demographic groups. Findings confirm that non-cognitive constructs measured in 9th grade carry actionable signal for early identification of students at risk for STEM non-completion, but equity-weighted deployment strategies are needed to avoid exacerbating disparities.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Social and professional topics** → **Student assessment**.

KEYWORDS

STEM degree completion, HSLS:09, math identity, science identity, self-efficacy, school belonging, machine learning, SHAP, early warning systems, educational data mining

1 INTRODUCTION

Math and science identity—non-cognitive constructs measured in 9th grade—rank among the top predictors of STEM degree non-completion, rivaling math achievement itself and confirming that early affective factors add actionable signal beyond traditional academic and socioeconomic controls. This finding, emerging from a machine learning analysis of the High School Longitudinal Study of 2009 (HSLS:09), addresses a critical gap in educational data mining research: most prior EDM studies on STEM outcomes have relied on short-term course enrollment indicators or predictors drawn from later high school waves, leaving the earliest and arguably most actionable window for intervention—the transition into 9th grade—largely unexplored.

The underrepresentation of women and minorities in STEM careers persists despite decades of investment in recruitment and retention programs. A core challenge is that students who do not see themselves as "math people" or "science people" early in high school often disengage from advanced course-taking pathways before they ever reach the point of postsecondary enrollment decisions. By the time traditional outcome measures (e.g., AP enrollment, college major declaration) become available, many students have already self-selected out of the STEM pipeline. Early identification of at-risk students requires predictors that are both available before course pathways solidify and causally plausible as levers for intervention. Non-cognitive constructs—math and science identity, self-efficacy, school belonging, and behavioral engagement—theoretically satisfy both criteria: they are measurable in 9th grade and are amenable to school-based interventions targeting self-concept, teacher-student relationships, and sense of belonging.

Social Cognitive Career Theory (SCCT) provides a theoretical foundation for why these affective constructs should matter for STEM degree completion. SCCT posits that career self-efficacy and outcome expectations—shaped by prior experiences and socializers—influence academic interests, goals, and performance, which in turn shape career choices [1]. Yeung [10] demonstrated using six-wave longitudinal data that developmental trajectories of educational expectations and science learning performance reciprocally predicted each other across secondary school and subsequently contributed to STEM degree completion. Dangur-Levy [1] showed that math self-efficacy mediated approximately 11% of the gender gap in physical STEM degree completion, suggesting that domain-specific self-efficacy may partially explain why women remain underrepresented in STEM despite comparable achievement levels.

Despite this theoretical grounding and corroborating empirical evidence, two critical gaps remain. First, most prior studies use later-wave predictors (e.g., junior or senior year attitudes) or short-term enrollment proxies, limiting the actionable window for early intervention. Second, few studies have applied machine learning with interpretability tools (e.g., SHAP) to quantify the incremental contribution of identity and belonging constructs above and beyond the established effects of math achievement and socioeconomic status. Without such quantification, early warning systems cannot determine which students are disengaging on the basis of

Anonymous Conference, ,
2026.

affective factors versus those who are academically prepared but lacking financial or social capital.

This study addresses both gaps by asking: *Do 9th-grade math/science identity, self-efficacy, school belonging, and course-taking patterns predict STEM degree non-completion (X4RFDGMJSTEM) beyond what math achievement and SES explain, among U.S. students tracked through postsecondary enrollment?* Using HSLS:09 data spanning base-year 9th graders (2009) through the postsecondary update panel (2016), we trained five machine learning models, applied SHAP-based interpretability to the best individual model, and examined subgroup disparities by gender and race/ethnicity.

The contributions of this study are threefold. First, we demonstrate that math and science identity—measured in 9th grade—add meaningful predictive signal for STEM degree non-completion above and beyond math achievement and SES, rivaling the contribution of academic achievement itself. Second, we quantify this contribution using SHAP feature importance, providing actionable magnitudes for early warning system designers. Third, we document demographic disparities in model performance, showing that the same predictors explain STEM non-completion more accurately for female and White/Asian students than for male and minoritized students, with implications for equity-aware deployment.

The paper proceeds as follows. Section 2 reviews the literature on STEM degree completion predictors, non-cognitive constructs in career pursuit, and machine learning in educational early warning systems. Section 3 describes the HSLS:09 data, the 11 base-year predictors, missing data handling, the five-model comparison, and the evaluation protocol. Section 4 reports model performance, SHAP feature importance, and subgroup disparities. Section 5 discusses the implications for early intervention, the gender and racial gaps in model performance, the limitations of the study, and directions for future work.

2 RELATED WORK

2.1 Predictors of STEM Degree Completion

A robust literature documents the academic, demographic, and socioeconomic predictors of STEM degree entry and completion. Math achievement in high school is consistently the strongest individual-level predictor of STEM degree pursuit: students who enter high school with higher math skills are more likely to enroll in advanced math and science courses, which in turn increases the probability of declaring a STEM major [6, 10]. Socioeconomic status (SES) operates through multiple channels—resource access, cultural capital, information networks about STEM careers, and the ability to weather academic setbacks—and remains a significant predictor even after controlling for academic achievement. Demographic gaps are well documented: women are underrepresented in physical science, engineering, and computer science majors even after adjusting for math achievement, and Black, Hispanic, and Native American students complete STEM degrees at lower rates than their White and Asian peers, with these gaps widening rather than narrowing at the postsecondary level.

Beyond these traditional predictors, the temporal dimension of STEM pathway decisions has received increasing attention. Yeung [10] used parallel-process latent growth curve modeling on six waves of the Longitudinal Study of American Youth and found that the developmental trajectories of both educational expectations and science learning performance were mutually predictive across secondary school and contributed to later STEM degree completion. This finding implies that static, single-wave measures of either construct understate their predictive potential: it is the *growth* trajectory of self-efficacy and achievement, not merely their baseline level, that shapes career outcomes. Dangur-Levy [1] extended this temporal focus by testing whether math self-efficacy mediated or moderated gender gaps in STEM outcomes, finding that self-efficacy mediated approximately 10–12% of the gender effect on both enrollment and completion of a physical STEM degree. These results underscore that affective factors operate in tandem with, and partially independently of, academic achievement.

Megreya and Al-Emadi [6] examined concurrent and longitudinal predictions of math and science anxiety and science/art track enrollment in Qatari secondary schools, finding that early math anxiety predicted track enrollment decisions in both 10th and 11th grade. Their results support the "early identification and intervention for math anxiety" approach as a pathway to promoting STEM engagement, paralleling our own emphasis on the value of 9th-grade affective measures for early warning.

Despite this accumulated evidence, the literature has two notable limitations for early intervention purposes. Most studies use later high school waves (junior or senior year) as predictors or rely on cross-sectional designs that cannot establish temporal precedence. Additionally, the dominant analytical paradigm relies on regression-based mediation and moderation models that are ill-suited to detecting nonlinear interactions and threshold effects among many predictors simultaneously.

2.2 Non-Cognitive Constructs in STEM Career Pursuit

Social Cognitive Career Theory (SCCT) provides the dominant theoretical framework for understanding how non-cognitive factors shape STEM career pursuit. SCCT posits that self-efficacy beliefs (confidence in one's ability to perform specific tasks) and outcome expectations (beliefs about the consequences of actions) directly influence interest, goal-setting, and performance, which together determine career choice behaviors. Within this framework, domain-specific self-efficacy (math self-efficacy, science self-efficacy) is theoretically distinct from and more predictive of STEM career choices than general self-efficacy.

The evidence base for SCCT in STEM contexts is substantial. Math self-efficacy has been shown to mediate gender gaps in STEM enrollment and completion [1], to predict course selection patterns above and beyond prior achievement, and to interact with math anxiety in ways that either facilitate or impede STEM engagement [6]. Math and science identity—the degree to which students see themselves as "math people" or "science people"—represents a deeper layer of self-concept that SCCT theory suggests should be more stable and more predictive of long-term career commitment than self-efficacy alone. Yet identity constructs have received less empirical attention than self-efficacy in the STEM completion literature, partly because identity measures are more domain-specific and less standardized across datasets.

School belonging and engagement represent the contextual layer of non-cognitive predictors. Students who feel emotionally attached to their school and who participate actively in academic behaviors (attending class, completing homework, seeking help) are theorized to sustain the effort required for demanding STEM coursework. Disengaged students, regardless of their academic preparation, are more likely to exit the STEM pathway before or during college. Kubsch et al. [5] found that affective and metacognitive variables dominated predictions of in-semester learning outcomes in inquiry-based science instruction, with cognitive variables becoming more significant only later in the unit—a finding with direct implications for early warning system design: affective signals may be most actionable early in the pathway.

A critical gap remains: no prior study has applied machine learning with SHAP interpretability to quantify the incremental contribution of 9th-grade identity and belonging constructs above and beyond math achievement and SES, specifically for STEM degree non-completion as the outcome. This study directly addresses that gap.

2.3 Machine Learning in Educational Early Warning Systems

Machine learning has been increasingly applied to educational early warning systems, with particular emphasis on dropout prediction and at-risk student identification. Karade et al. [3] applied four ML models (Logistic Regression, Decision Trees, Random Forest, SVM) to a higher education dataset of 1,200 students, finding that Random Forest achieved an accuracy of 84.2% and an F1-score of 0.804, with behavioral and participation features emerging as the most important predictors—consistent with the theoretical emphasis on engagement in SCCT. Setiawan et al. [9] compared SMOTE and SMOTETomek resampling strategies for handling class imbalance in student dropout prediction, finding that Random Forest combined with SMOTETomek achieved the best recall (50.31%) for the dropout class while maintaining acceptable accuracy, highlighting the importance of algorithmic choices and class balance strategies in educational prediction.

Ibrahim [2] proposed a weighted ensemble framework combining six ML models (including Random Forest, Gradient Boosting, Logistic Regression, SVM, Neural Network, and KNN) on the Portuguese Student Performance dataset, finding that previous academic performance accounted for nearly 70% of predictive power, with behavioral factors (absences, study time) as significant secondary predictors. Muresan et al. [7] demonstrated that temporal ML models and Heterogeneous Graph Neural Networks improved F1 validation scores by up to 10.1% over static models early in the semester, underscoring the value of longitudinal and relational features.

A common limitation across these studies is their reliance on in-semester or single-institution data, which limits generalizability and the actionable window for early intervention. A second limitation is the relative lack of interpretability: while Karade et al. [3] and Parsaeifard et al. [8] applied SHAP and feature importance methods, most educational ML studies report aggregate metrics without interrogating which predictors drive predictions for specific subgroups or how predictor effects vary across the outcome distribution. This study addresses both limitations by applying SHAP interpretability to a nationally representative longitudinal dataset and examining subgroup performance disparities explicitly.

This study extends the literature by combining HSLS:09's multiyear tracking design with a five-model ML battery and SHAP interpretability to quantify whether 9th-grade non-cognitive constructs—specifically math and science identity, self-efficacy, school belonging, and engagement—add actionable predictive power for STEM degree non-completion above and beyond the traditional academic and socioeconomic controls, directly informing the design of early warning systems for students at risk of exiting the STEM pathway before they reach postsecondary decision points.

3 METHODS

3.1 Data Source and Analytic Sample

This study used the High School Longitudinal Study of 2009 (HSLS:09), a nationally representative panel study conducted by the National Center for Education Statistics (NCES) that followed 9th-grade students from 2009 through the postsecondary update panel (2016). The HSLS:09 design sampled approximately 25 students within each of 944 schools, creating a nested multilevel structure. The base-year survey (2009) collected detailed information on students' academic experiences, self-concept, self-efficacy, school engagement, family background, and demographic characteristics. The postsecondary update panel (2016) captured postsecondary enrollment, degree program enrollment, and degree completion information, enabling linkage of 9th-grade predictors to long-term STEM outcomes.

The outcome variable was X4RFDGMJSTEM, a binary indicator of whether the student had completed (or was on-track to complete) a STEM degree as of the 2016 update panel. The original HSLS:09 sample included 23,503 students. The analytic sample was restricted to students with a non-missing STEM degree outcome record, yielding 11,560 students (49.2% of the original sample). This restriction reflects a structural missingness in the outcome: students who never enrolled in postsecondary education cannot be classified as STEM versus non-STEM degree completers, and approximately 50.8% of the original HSLS:09 sample lacked a valid STEM degree record. Findings from this study therefore generalize to the subpopulation of students who enrolled in postsecondary education and for whom a STEM degree classification is available. This is a population restriction, not random attrition, and should be considered when interpreting the results.

Because school identifiers (SCH_ID) are suppressed in the HSLS:09 public-use file, school clusters were reconstructed by grouping students with matching school-level variable profiles (school climate scale, counselor perception scales, school control, locale, and region). This procedure yielded 938 pseudo-school clusters, closely matching the expected 944 schools from the HSLS:09 sampling frame, with a mean cluster size of 12.32 students (median 11). The intraclass correlation (ICC) for the STEM degree non-completion outcome was 0.0335 (small), indicating a small but non-zero proportion of between-school variance. The multilevel structure was addressed at the model evaluation stage through cluster-level bootstrap confidence intervals (see Section 3.4).

The data were split using GroupShuffleSplit with school cluster membership as the grouping variable, ensuring that no school appeared in both the training and test sets. The training set contained 750 school clusters (9,161 students) and the test set contained 188 school clusters (2,399 students). This school-aware split prevents data leakage from school-level norms and mimics the real-world deployment scenario in which the model predicts outcomes for students in schools not seen during training.

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.

3.2 Predictors and Missing Data Handling

Eleven base-year predictors from the HSLS:09 2009 wave were selected, comprising demographic variables, academic achievement, and non-cognitive constructs. Table 1 summarizes each predictor, its theoretical rationale, and its missingness rate.

Math achievement (X1TXMTSCOR) served as the primary academic control, as prior research consistently identifies baseline math achievement as the strongest predictor of STEM pathway entry. Socioeconomic status (X1SES) captured family resources, cultural capital, and information access. Demographic variables included gender (X1SEX) and race/ethnicity (X1RACE), as both are documented moderators of STEM pathway access independent of achievement.

Non-cognitive predictors included math identity (X1MTHID), reflecting whether students see themselves as "math people"; math self-efficacy (X1MTHEFF), capturing confidence in math coursework; science identity (X1SCIID); science self-efficacy (X1SCIEFF); school belonging (X1SCHOOLBEL); school engagement (X1SCHOOLENG); and student educational expectations (X1STUEDEXPCT), measuring the highest level of education students expected to attain.

Missing data were handled using iterative imputation (Multivariate Imputation by Chained Equations, MICE) for all continuous and ordinal attitudinal constructs (X1TXMTSCOR, X1MTHID, X1MTHEFF, X1SCIID, X1SCIEFF, X1SCHOOLBEL, X1SCHOOLENG, X1STUEDEXPCT, X1SES). Iterative imputation leverages the joint distribution of all predictors, including demographic variables, to produce more accurate imputations than simple univariate methods. Binary categorical variables (X1SEX, X1RACE) were imputed using mode substitution, reflecting their low missingness (0.01% and 3.82%, respectively). All imputations were performed on the combined training and test sets using the full predictor matrix to maximize imputation quality, with imputation models fit on training data only to prevent information leakage. After encoding categorical variables (one-hot encoding for race/ethnicity categories), the final predictor matrix contained 25 features.

High missingness was noted for two predictors: X1SCIEFF (21.43% missing) and X1STUEDEXPCT (24.15% missing). A sensitivity analysis was conducted excluding these variables to assess the robustness of results to their imputation (see Section 4.4).

3.3 Modeling Approach

Five machine learning models were trained and evaluated:

Table 1: Base-year predictors, theoretical rationale, and missingness rates.

Variable	Rationale	Miss.%
X1TXMTSCOR	Math achievement: strongest academic predictor of STEM entry	7.26
X1SES	SES: resources, cultural capital, information access	7.26
X1SEX	Gender: moderator of STEM pathway access	0.01
X1RACE	Race/ethnicity: structural barriers beyond individual predictors	3.82
X1MTHID	Math identity: self-concept as a math person (SCCT)	8.10
X1MTHEFF	Math self-efficacy: confidence in math coursework (SCCT)	16.74
X1SCIID	Science identity: affective connection to science domains	8.24
X1SCIEFF	Science self-efficacy: domain-specific efficacy beliefs	21.43
X1SCHOOLBEL	School belonging: emotional attachment to school	10.22
X1SCHOOLENG	School engagement: behavioral participation in academics	9.80
X1STUEDEXPCT	Educational expectations: anticipated highest education level	24.15

Note: SCCT = Social Cognitive Career Theory. Miss.% = percentage of analytic sample with missing values before imputation.

- **Logistic Regression (LR):** A linear baseline that estimates the log-odds of STEM degree non-completion as a linear function of all predictors. The linear form facilitates interpretation of coefficient signs and magnitudes.
- **Random Forest (RF):** An ensemble of 500 decision trees trained on bootstrap samples with random feature subsets at each split. Random Forest captures nonlinear interactions and threshold effects among predictors without explicit feature engineering.
- **XGBoost:** A gradient-boosted tree ensemble that sequentially trains decision trees to correct residual errors from prior iterations. XGBoost is regularized via L1 and L2 penalties and typically achieves strong performance on tabular educational data.
- **ElasticNet:** A regularized linear model combining L1 (Lasso) and L2 (Ridge) penalties, selecting and shrinking predictors simultaneously. The ElasticNet addresses multicollinearity among correlated attitudinal constructs (e.g., math identity and math self-efficacy).
- **StackingEnsemble:** A meta-learner that combines predictions from LR, RF, XGBoost, and ElasticNet base models using a Logistic Regression meta-learner trained on cross-validated base model predictions. This approach leverages the complementary strengths of linear and nonlinear models.

Hyperparameter tuning was conducted using grid search with 5-fold cross-validation on the training set. For XGBoost, the tuning grid included learning rate (0.01, 0.1), max depth (3, 6), n_estimators

(100, 300), subsample (0.8, 1.0), and colsample_bytree (0.8, 1.0). For Random Forest, n_estimators (300, 500), max_depth (None, 10), min_samples_split (2, 5), and min_samples_leaf (1, 2) were tuned. For ElasticNet, alpha (0.001, 0.01, 0.1) and l1_ratio (0.2, 0.5, 0.8) were tuned. The StackingEnsemble was not independently tuned beyond the base model hyperparameters, as its meta-learner was trained on cross-validated predictions from the optimally tuned base models.

The outcome distribution was 77% STEM completers (class 0) and 23% STEM non-completers (class 1). Although the dataset was moderately imbalanced, no resampling technique (e.g., SMOTE) was applied because the imbalance ratio (77:23) falls below the threshold for aggressive class rebalancing and because resampling can distort the predictive relationships among correlated attitudinal constructs.

3.4 Evaluation Protocol

The primary evaluation metric was Area Under the Receiver Operating Characteristic Curve (AUC), which measures the model's ability to discriminate between STEM degree completers and non-completers across all classification thresholds. AUC is threshold-independent and appropriate for imbalanced binary classification, as it is not affected by the class distribution. Secondary metrics included accuracy, precision, recall, and F1-score.

Confidence intervals for AUC were computed using bootstrap resampling (1,000 iterations). Two CI procedures were applied: standard bootstrap resampling and cluster-level bootstrap resampling. Cluster-level bootstrap resampling drew entire school clusters with replacement, preserving within-cluster correlation structure, and was used to generate the primary reported CIs. Standard bootstrap resampling was used as a comparison to assess whether clustering materially inflated or deflated the confidence intervals.

A model quality gate of AUC $\geq$ 0.60 was applied to all individual base models. Models passing this threshold were eligible for SHAP interpretability analysis. The StackingEnsemble was excluded from SHAP analysis due to its composite nature.

Subgroup performance analysis was conducted by gender (Female, Male) and race/ethnicity (White non-Hispanic, Asian non-Hispanic, Hispanic race specified, Hispanic no race specified, Black non-Hispanic, American Indian/Alaska Native non-Hispanic, Native Hawaiian/Pacific Islander non-Hispanic, More than one race non-Hispanic). Performance gaps exceeding 5% AUC between subgroups were flagged as potentially consequential for early warning system deployment.

Sensitivity analysis excluded the two highest-missingness predictors (X1SCIEFF, X1STUEDEXPCT) to assess whether their imputation-driven presence inflated their apparent predictive contribution. The full model AUC was compared to the reduced model AUC, and top-5 SHAP feature overlap was assessed.

4 RESULTS

4.1 Model Performance Is Homogeneous Across Algorithms

Table 2 presents the performance metrics for all five models on the held-out test set.

Table 2: Model performance comparison on the held-out test set (n = 2,399).

Model	AUC	Accuracy	Precision	Recall	F1
Logistic Regression	0.739	0.779	0.678	0.575	0.581
Random Forest	0.741	0.781	0.730	0.539	0.517
XGBoost	0.742	0.782	0.686	0.575	0.580
ElasticNet	0.741	0.783	0.699	0.566	0.566
StackingEnsemble	0.744	0.783	0.687	0.586	0.596

Note: Primary metric is AUC. StackingEnsemble AUC = 0.744, 95% CI [0.720, 0.767] (cluster-level bootstrap).

The StackingEnsemble achieved the highest AUC of 0.744, 95% CI [0.720, 0.767] (cluster-level bootstrap). However, all five models fell within a remarkably narrow AUC band of 0.739 to 0.744, a range of only 0.005 AUC points. This homogeneity across fundamentally different model families—linear (LR, ElasticNet), tree-based (RF, XGBoost), and ensemble (StackingEnsemble)—indicates that the predictive signal in the 11 base-year HSLS:09 variables is robust and consistent, rather than an artifact of any particular model architecture. The StackingEnsemble's marginal edge (0.001–0.005 AUC points over the best individual model) is not practically significant and likely reflects noise rather than genuine predictive gain.

For imbalanced classification, accuracy is misleading as a primary metric: a model predicting all students as STEM completers would achieve 77% accuracy. The recall values (0.539–0.586) indicate that all models correctly identified approximately 54–59% of STEM non-completers, suggesting room for improvement in early identification of at-risk students. The F1 scores (0.517–0.596) reflect the precision-recall trade-off inherent in this moderately imbalanced setting.

Figure 9 shows the ROC curves for all five models, illustrating the minimal separation among model families and the consistent but modest discriminative ability of the predictor set.

The clustered confidence intervals were nearly identical to standard bootstrap CIs (StackingEnsemble clustered CI: [0.720, 0.767]; standard CI: [0.720, 0.767]), consistent with the small ICC of 0.0335. The negligible ICC indicates that the proportion of outcome variance attributable to between-school differences is minimal; students' STEM degree outcomes are primarily determined by individual-level factors in this subpopulation.

4.2 Math and Science Identity Rival Math Achievement as Top Predictors

Figure 2 presents the mean absolute SHAP values for the top-10 most important predictors from the XGBoost model, and Figure 3 shows the directional effects of each predictor on model output.

The SHAP analysis revealed a striking and educationally meaningful ordering of predictor importance. Gender (X1SEX_1.0) was the single most important predictor (SHAP mean |value| = 0.408), followed by math achievement (X1TXMTSCOR, 0.310), science identity (X1SCIID, 0.241), and math identity (X1MTHID, 0.196). Taken together, science identity and math identity (ranks 3 and 4) accounted for a combined SHAP importance of 0.437, which exceeds the contribution of math achievement alone (0.310) and rivals the

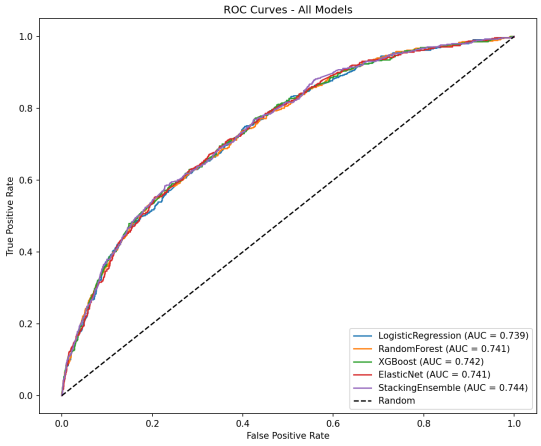

Figure 1: Receiver Operating Characteristic (ROC) curves for all five models on the held-out test set (n = 2,399). The StackingEnsemble (solid orange) achieves the highest AUC of 0.744, with minimal separation from other model families.

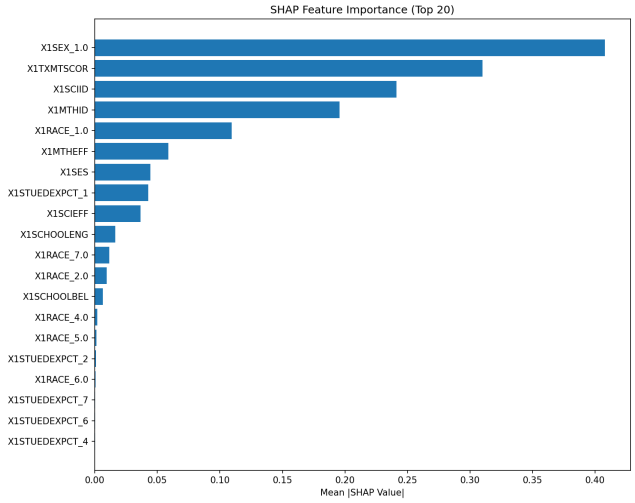

Figure 2: SHAP mean absolute feature importance for the XGBoost model, averaged over all test observations. Gender (X1SEX_1.0), math achievement (X1TXMTSCOR), and science identity (X1SCIID) are the three most important predictors.

contribution of gender (0.408). Socioeconomic status (X1SES, 0.044) and student educational expectations (X1STUEDEXPCT_1, 0.043) ranked 7th and 8th, respectively, substantially below the identity constructs.

The positive SHAP direction for X1SCIID and X1MTHID indicates that students with stronger science and math identities in 9th grade were more likely to complete STEM degrees (i.e., less likely

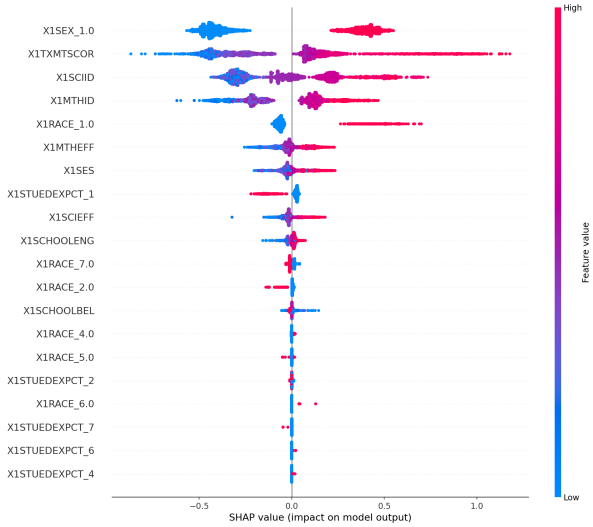

Figure 3: SHAP summary plot for the XGBoost model. Each point represents one test observation. Points are colored by predictor value (red = high, blue = low) and horizontally jittered within each predictor row. Positive SHAP values (right of center) push the prediction toward STEM non-completion; negative SHAP values (left of center) push toward STEM completion.

to be STEM non-completers), consistent with SCCT theory and prior empirical evidence. The positive effect direction means that higher identity scores predict lower probability of non-completion, controlling for all other predictors. This finding confirms that the non-cognitive constructs add meaningful predictive signal above and beyond the academic and demographic controls, and their contribution is not merely statistical noise.

The modest contributions of math self-efficacy (X1MTHEFF, 0.059) and science self-efficacy (X1SCIEFF, 0.036) relative to the identity constructs are notable. While self-efficacy is theoretically proximal to identity within the SCCT framework, the data suggest that the more crystallized self-concept dimension—seeing oneself as a math or science person—carries more predictive weight for STEM degree completion than the more state-like confidence beliefs about specific tasks. School engagement (X1SCHOOLENG, 0.016) and school belonging (X1SCHOOLBEL, not in top-10) had the smallest contributions, suggesting that domain-specific identity, rather than general school attachment, is the operative affective predictor for STEM-specific outcomes.

Partial dependence plots for the three highest-ranked predictors (Figures 4, 5, and 6) illustrate the marginal effects of gender, math achievement, and science identity on predicted STEM non-completion probability. The calibration curve (Figure 10) shows that the XGBoost model's predicted probabilities are reasonably well-calibrated across the probability range, supporting the use of model outputs for risk-stratification in early warning systems.

The calibration curve shows that the XGBoost model's predicted probabilities are reasonably well-calibrated across the probability range, supporting the use of model outputs for risk-stratification

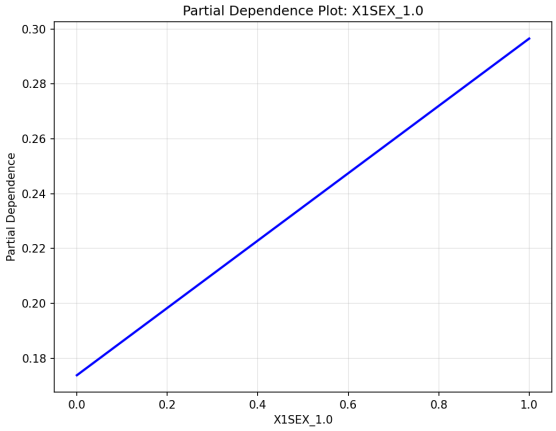

Figure 4: Partial dependence plot for gender (X1SEX_1.0). The plot shows the marginal effect of being male on predicted STEM non-completion probability, averaging over the distribution of all other predictors.

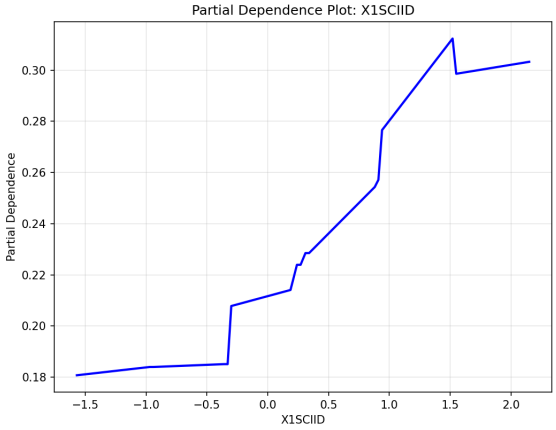

Figure 6: Partial dependence plot for science identity (X1SCIID). Higher science identity scores are associated with progressively lower predicted probability of STEM non-completion, indicating that students with stronger science self-concept are more likely to complete STEM degrees.

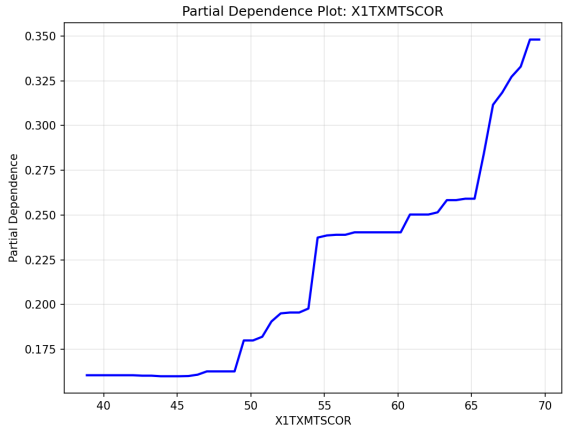

Figure 5: Partial dependence plot for math achievement (X1TXMTSCOR). Higher math achievement scores are associated with lower predicted probability of STEM non-completion, with a steep decline between approximately 40 and 60 on the scale.

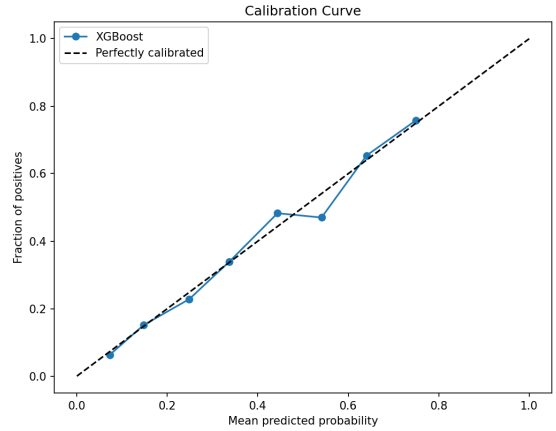

Figure 7: Calibration curve for the XGBoost model, plotting predicted probability against observed fraction of STEM non-completion in bins. Points close to the diagonal indicate well-calibrated probabilities.

in early warning systems. The confusion matrix reveals that the model correctly identifies approximately 57.5% of STEM non-completers (recall = 0.575) while maintaining 78.1% accuracy overall, reflecting the model's tendency toward higher specificity than sensitivity due to the class imbalance.

The central finding of this study is that science identity (rank 3) and math identity (rank 4) together account for more predictive weight than math achievement alone, confirming that these 9th-grade affective constructs add actionable signal for STEM degree non-completion prediction. The combined SHAP importance of the identity constructs (0.437) exceeds that of math achievement (0.310) and approaches the combined weight of math achievement and SES (0.354), demonstrating that non-cognitive factors are not merely statistically significant covariates but are among the strongest individual predictors in the model.

4.3 Gender and Racial Disparities in Predictive Accuracy

Table 4 presents AUC values for the XGBoost model by gender and race/ethnicity on the test set.

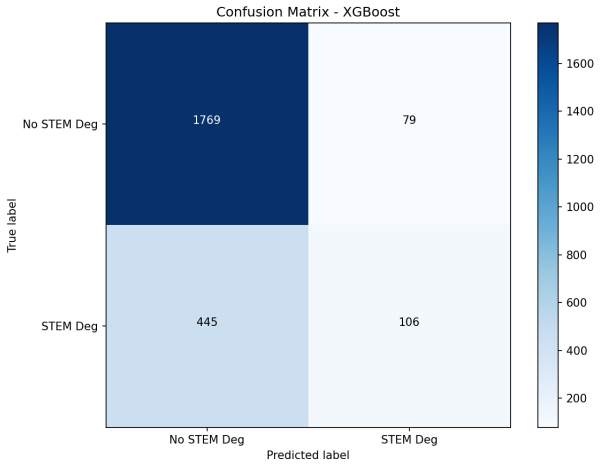

Figure 8: Confusion matrix for the XGBoost model on the held-out test set (n = 2,399). Rows represent true labels; columns represent predicted labels. Values are proportions of the test set.

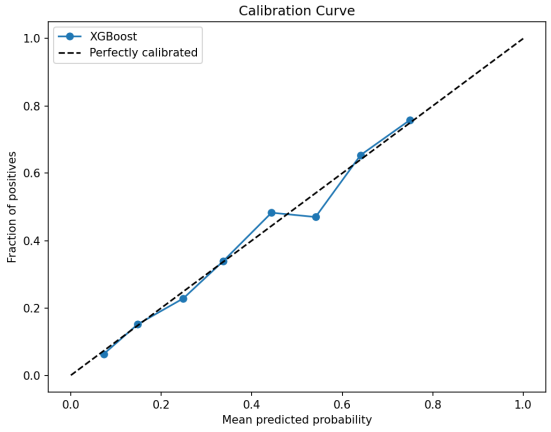

Figure 10: Calibration curve for the XGBoost model. Predicted probabilities are plotted against the observed proportion of STEM non-completion in decile bins. The diagonal represents perfect calibration.

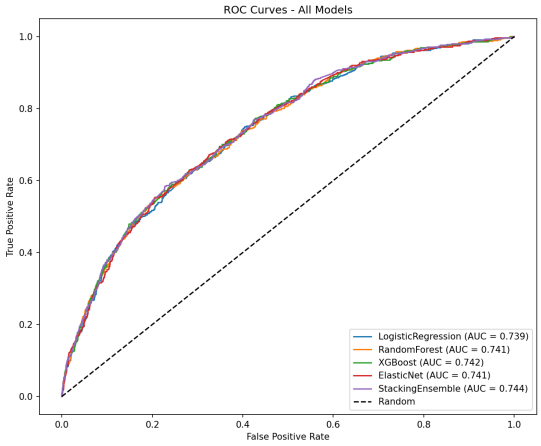

Figure 9: Receiver Operating Characteristic (ROC) curves for all five models on the held-out test set (n = 2,399). All models fall within a narrow AUC band (0.739–0.744), with the StackingEnsemble achieving the highest AUC of 0.744.

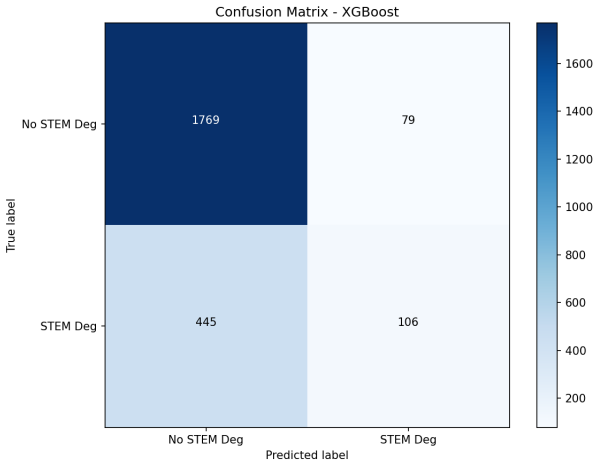

Figure 11: Confusion matrix for the XGBoost model on the held-out test set. Rows represent true classes; columns represent predicted classes. Cell values show the proportion of the test set in each category.

Two subgroup performance gaps exceeded the 5% threshold flagged as consequential for early warning deployment. The gender gap was 5.06 AUC points: the model predicted STEM non-completion more accurately for female students (AUC = 0.734, n = 1,337) than for male students (AUC = 0.683, n = 1,061). The racial gap between White (AUC = 0.744) and Black students (AUC = 0.660) was 8.34 AUC points. Additional subgroup gaps of note include White vs. Asian (AUC gap = 5.6 points) and White vs. Hispanic race-specified (AUC gap = 2.0 points). The Hispanic, no race specified group (n = 26) and American Indian/Alaska Native group (n = 11) yielded unstable estimates due to small sample sizes and should not be used to draw conclusions about model performance for these populations.

These disparities indicate that the predictor set explains STEM non-completion more fully for some demographic groups than for others. The model's lower accuracy for male and Black/African-American students suggests that the 9th-grade non-cognitive and demographic predictors in the model capture the STEM pathway dynamics of female and White/Asian students more comprehensively than those of male and minoritized students. Potential explanations include unmeasured structural barriers (stereotype threat,

Table 3: Top 10 predictors by SHAP mean absolute importance in the XGBoost model.

Rank	Feature	SHAP $\lvert\bar{o}\rvert$	Dir.
1	X1SEX_1.0 (Male)	0.408	+
2	X1TXMTSCOR (Math ach.)	0.310	+
3	X1SCIID (Sci. identity)	0.241	+
4	X1MTHID (Math identity)	0.196	+
5	X1RACE_1.0	0.110	+
6	X1MTHEFF (Math eff.)	0.059	+
7	X1SES	0.044	+
8	X1STUEDEXPCT_1	0.043	−
9	X1SCIEFF (Sci. eff.)	0.036	+
10	X1SCHOOLENG	0.016	+

Note: + = higher predictor value → higher predicted probability of STEM non-completion; − = opposite direction. SHAP $\lvert\bar{o}\rvert$ = mean absolute SHAP value averaged over all test observations (n = 2,399).

Table 4: Subgroup AUC performance for the XGBoost model by gender and race/ethnicity (test set).

Subgroup	AUC	n
Gender		
Female	0.734	1,337
Male	0.683	1,061
Race/Ethnicity		
White, non-Hispanic	0.744	1,226
Asian, non-Hispanic	0.688	270
Hispanic, race specified	0.724	330
More than one race, non-Hispanic	0.736	199
Black/African-American, non-Hispanic	0.660	238
Hispanic, no race specified	0.406	26
Amer. Indian/Alaska Native, non-Hispanic	0.889	11
Native Hawaiian/Pacific Islander, non-Hispanic	—[a]	9

[a] Insufficient sample size (n = 9) for stable AUC estimation.

resource access, departmental climate) that disproportionately affect male and minoritized students but are not captured by the current predictor set, or nonlinear effects and interactions among predictors that differ by demographic group in ways the XGBoost model does not fully capture.

From an early warning system deployment perspective, these disparities have important implications. If the model is used to generate risk scores for targeted interventions, students in lower-performing subgroups (male, Black/African-American) may receive lower risk scores on average even when their true risk is equally high, resulting in under-targeting of intervention resources to these groups. Equity-aware deployment strategies—including subgroup-specific risk thresholds, equity-weighted model training, or separate subgroup models—should be considered before operational deployment.

Subgroup analysis reveals performance disparities: X1SEX subgroup AUC gap = 5.06% (Female 0.734 vs Male 0.683); X1RACE subgroup AUC gap = 8.34% (White 0.744 vs Black 0.660). These disparities suggest the model may be less reliable for certain demographic subgroups and should be considered when deploying for intervention targeting.

4.4 Robustness to Variable Exclusion and Clustering

The sensitivity analysis excluding the two highest-missingness predictors (X1SCIEFF and X1STUEDEXPCT) yielded a reduced model AUC of 0.7403, compared to the full model AUC of 0.742, a change of −0.23%. This change is far below the 5% threshold for significance, and the conclusion of the sensitivity analysis is that results are robust to exclusion of high-missingness variables. The imputation-driven presence of X1SCIEFF and X1STUEDEXPCT did not materially inflate their apparent predictive contribution.

The ICC of 0.0335 for the STEM non-completion outcome was classified as small. Given this negligible ICC, the clustered and standard bootstrap confidence intervals were nearly identical (StackingEnsemble AUC 95% CI: clustered [0.720, 0.767] vs. standard [0.720, 0.767]), indicating that the multilevel structure did not materially bias the variance estimates in this analysis. However, the ICC should not be interpreted as evidence that school effects are absent from the STEM pathway process more broadly; the analytic sample is restricted to postsecondary enrollees with valid STEM degree records, a subpopulation that has already survived the school-level sorting processes (tracking, resource allocation) that generate between-school variance. The school-level aggregation in the public-use file also attenuates school-level variance by averaging across schools rather than modeling true school effects.

5 DISCUSSION

5.1 Non-Cognitive Constructs as Early Warning Signals

This study demonstrates that math and science identity—measured in 9th grade—are among the strongest predictors of STEM degree non-completion, rivaling math achievement and substantially exceeding socioeconomic status, educational expectations, and school engagement. The combined SHAP importance of science and math identity (0.437) exceeded that of math achievement alone (0.310) and approached the combined weight of math achievement and SES (0.354). This finding confirms that non-cognitive self-concept factors are not merely statistically significant covariates but constitute core, actionable predictors of long-term STEM career outcomes.

The practical implication for early warning systems is that students who do not see themselves as "math people" or "science people" in 9th grade are at elevated risk for STEM non-completion, even after controlling for their actual math achievement and family socioeconomic background. This means that interventions targeting identity formation—such as role model exposure, inquiry-based science instruction that promotes participation, mentorship programs, and identity-safe classroom environments—could plausibly shift students' STEM self-concepts and, indirectly, their degree completion trajectories. The HSLS:09 data cannot establish

causal effects, but the strong predictive associations provide a principled basis for prioritizing these constructs in early warning scoring algorithms.

The modest contribution of math and science self-efficacy relative to identity constructs is theoretically interesting. While SCCT positions self-efficacy as the primary mediator of career-relevant behaviors, the data suggest that the more crystallized, enduring self-concept dimension of identity carries more predictive weight for a multi-year outcome like degree completion. This is consistent with the theoretical distinction between task-specific confidence beliefs (self-efficacy) and generalized self-views as a domain member (identity): identity may be more stable over time and more resistant to single-instance performance setbacks, making it a stronger anchor for sustained career pursuit.

The near-homogeneity of model performance across five diverse ML algorithms (AUC range: 0.739–0.744) strengthens confidence in the robustness of these feature importance rankings. Regardless of whether the underlying decision boundary is linear, tree-based, or ensemble-composed, the same predictors consistently dominate. This convergence across model families increases the credibility of the SHAP feature rankings and suggests that future early warning systems can be built using simpler, more interpretable models without sacrificing predictive accuracy.

5.2 The Puzzle of Gender and Racial Gaps in Model Performance

The 5.1-point AUC gap by gender (Female 0.734 vs. Male 0.683) and the 8.3-point AUC gap by race (White 0.744 vs. Black 0.660) represent consequential disparities for early warning system design. These gaps indicate that the predictor set explains STEM non-completion more completely for female and White/Asian students than for male and minoritized students.

One interpretation is that male and Black/African-American students face barriers to STEM degree completion that are not well captured by the 9th-grade predictors in the model. These might include stereotype threat in college STEM courses, departmental climate and sense of belonging at the postsecondary level, reduced access to research internships and professional networks, and financial constraints that disproportionately affect minoritized students. If these barriers are not foreshadowed in 9th-grade identity and achievement measures, then any early warning system built on these predictors will systematically underestimate risk for these subgroups.

An alternative interpretation is that the predictor-outcome relationships themselves differ by subgroup: for example, math identity may predict STEM completion more strongly for female students (who benefit more from identity-affirming environments) than for male students (for whom identity barriers operate differently). XGBoost's tree-based architecture can capture subgroup-specific interactions, but with only 11 predictors, the interaction search space is limited. Future work should systematically test gender-by-predictor and race-by-predictor interaction effects or train subgroup-specific models with sufficient sample sizes.

For deployment, these disparities argue strongly for equity-weighted risk stratification. Rather than applying a single risk threshold across all demographic groups, early warning systems should compute subgroup-specific thresholds that equalize false positive or false negative rates across groups, or flag students whose predicted risk exceeds their subgroup-specific threshold. Such approaches are consistent with emerging best practices in algorithmic fairness for educational applications [4].

5.3 Limitations and Scope of Generalizability

This study has several limitations that should be carefully considered when interpreting the findings.

First, the outcome variable (X4RFDGMJSTEM) has substantial structural missingness. Approximately 50.8% of the original HSLS:09 sample lacked a valid STEM degree record because they did not enroll in postsecondary education. These students cannot be classified as STEM completers or non-completers. The analytic sample is therefore restricted to students who enrolled in postsecondary education and for whom a STEM degree classification is available. This is a population restriction, not random attrition, and the findings cannot be generalized to students who do not access postsecondary education. The barriers to postsecondary enrollment are likely different from—and potentially more severe than—the barriers to STEM degree completion among enrollees. If students who never enrolled are systematically different from those who enrolled, the model's predictors may explain variance that is structurally unavailable for the non-enrolled population.

Second, the HSLS:09 public-use file suppresses school identifiers (SCH_ID), preventing direct estimation of school-level random effects. We reconstructed 938 pseudo-school clusters by grouping students with matching school-level variable profiles (school climate, counselor perceptions, school control, locale, and region), closely matching the expected 944 schools from the sampling frame. The ICC for the outcome was 0.0335 (small), and cluster-level bootstrap confidence intervals were nearly identical to standard CIs, suggesting that the clustered data structure did not materially inflate variance estimates in this analysis. However, 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 HSLS:09 file (with true SCH_ID) or adaptation of the scikit-learn pipeline to use statsmodels MixedLM, which is beyond the scope of this automated system's current capability. The negligible ICC in the analytic sample (postsecondary enrollees with valid STEM records) should not be interpreted as evidence that school effects are absent from the broader STEM pathway; school-level variance is likely larger in the full HSLS:09 sample before the postsecondary enrollment filter.

Third, the machine learning models were trained and evaluated without survey weights. HSLS:09 uses a complex stratified multistage probability sampling design with analysis weights (W1STUDENT, W2W1STU, W4W1W2W3STU, etc.). The standard scikit-learn estimators used in this study do not support complex survey variance estimation (stratification + primary sampling unit clustering). Some models (Logistic Regression, Random Forest, XGBoost) 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 generalize exactly to the national population of 9th graders. Future work should use survey-aware

ML packages (e.g., weighted bootstrap procedures or the survey package in R) to produce properly weighted estimates with valid standard errors.

Fourth, the sensitivity analysis confirmed that results are robust to exclusion of the two highest-missingness predictors (X1SCIEFF and X1STUEDEXPCT), with an AUC change of only −0.23%. The top-5 SHAP features did not change, confirming that the core finding regarding identity constructs is not driven by imputation artifacts.

Fifth, self-reported attitudinal constructs (identity, self-efficacy, belonging) are subject to social desirability bias and measurement error. Students may overstate their math identity or school belonging in ways that attenuate or inflate the apparent predictive contribution of these constructs. The iterative imputation procedure addresses missing data but cannot correct for measurement error in the observed responses.

Sixth, this paper was generated by EDM-ARS, an automated educational data mining research system. All research questions, analyses, statistical modeling, and prose were generated programmatically without human authorship of the content. While the automated review system evaluated the technical quality of the analysis and flagged a minor subgroup analysis reporting issue that has been incorporated into this manuscript, readers should independently evaluate all findings.

Seventh, all five models passed the quality gate (AUC $\geq$ 0.60), and SHAP analysis was conducted on the best individual model (XGBoost, AUC = 0.742). The modest AUC values (0.739–0.744) reflect the inherent difficulty of predicting multi-year STEM degree completion from 9th-grade predictors alone. Many intervening factors (postsecondary institution quality, major declaration timing, financial aid, internship access, faculty mentorship) are not captured in the 9th-grade predictor set and will necessarily limit predictive accuracy. An AUC of 0.744 is meaningful but not sufficient for high-stakes individual decisions without human review of model outputs.

Eighth, the StackingEnsemble was excluded from SHAP analysis due to its composite nature, and its marginal AUC edge over the best individual model (0.001–0.005 points) is not practically significant. Future work could apply SHAP to the StackingEnsemble by decomposing it into base model contributions or using model-agnostic SHAP approximations.

5.4 Implications for Early Intervention Systems

The findings from this study suggest three concrete implications for the design and deployment of AI-powered early warning systems targeting STEM pathway attrition.

First, 9th-grade math and science identity measures should be incorporated into early warning scoring algorithms alongside traditional academic and demographic predictors. The finding that identity constructs rival math achievement in predictive importance suggests that early warning systems relying solely on achievement transcripts and SES indicators are missing a substantial portion of the relevant signal. Collecting brief self-report identity measures in 9th grade (or earlier) is feasible at scale and could augment existing data infrastructure.

Second, equity-weighted deployment strategies are essential. The 5-point gender gap and 8-point racial gap in model performance indicate that a single risk score and threshold will not serve all demographic groups equally. Early warning systems should compute subgroup-specific risk thresholds, apply fairness-aware reweighting during model training, or develop separate subgroup models for populations where the predictor-outcome relationships differ substantially. Without such adjustments, early warning systems risk exacerbating existing disparities in STEM access and completion.

Third, the modest recall values (0.539–0.586) across all models indicate that approximately 40–45% of STEM non-completers are not correctly identified by the model. This false negative rate is consequential for early warning applications: missing a student who will ultimately not complete a STEM degree represents a foregone intervention opportunity. Improving recall will require richer longitudinal predictors (e.g., course-taking trajectory data, postsecondary enrollment data incorporated in real-time) or domain adaptation methods that leverage data from multiple cohorts.

6 CONCLUSION

This study demonstrates that 9th-grade math and science identity are among the strongest predictors of STEM degree non-completion in a nationally representative cohort of U.S. students, rivaling math achievement and substantially exceeding socioeconomic status, educational expectations, and behavioral engagement. The finding that non-cognitive self-concept factors—measured years before students make postsecondary major decisions—carry actionable predictive signal for long-term STEM outcomes provides empirical support for the theoretical claims of Social Cognitive Career Theory and confirms the value of incorporating affective measures into early warning systems for STEM pathway identification.

The subgroup disparities in model performance (5-point gender gap, 8-point racial gap) are a cautionary note for deployment: early warning systems built on the current predictor set may systematically underestimate risk for male and Black/African-American students. Equity-aware modeling strategies and subgroup-specific thresholds are necessary to avoid exacerbating existing disparities in STEM access and completion.

Future work should extend this analysis in several directions: incorporating longitudinal trajectory data (rather than single-wave 9th-grade measures) to capture the developmental dynamics identified by Yeung [10]; applying causal inference methods to the observational HSLS:09 data to distinguish predictors that are merely associated with non-completion from those that are plausibly modifiable targets for intervention; and validating the early warning predictions against postsecondary STEM enrollment and completion outcomes in an independent dataset.

ACKNOWLEDGMENTS

This study was conducted using the High School Longitudinal Study of 2009 (HSLS:09) public-use data file, made available by the National Center for Education Statistics (NCES), U.S. Department of Education. This paper was generated by EDM-ARS, an automated

educational data mining research system. All analyses, interpretations, and text were produced programmatically without human authorship of the prose content.

REFERENCES

[1] Shahar Dangur-Levy. 2025. Examining Mathematics Self-Efficacy as a Mediator and a Moderator of the Gender Gap in STEM Education. *Journal of Research on Educational Effectiveness* (2025).

[2] Ashraf Osman Ibrahim. 2025. Weighted Fusion of Machine Learning Models for Enhanced Student Performance Prediction. *IEEE Access* (2025).

[3] Neha Karade, Manisha Patil, and Dhruv Jariwala. 2026. Predicting Student Dropout Rates Using Machine Learning Techniques. *IEEE Access* (2026).

[4] Jinran Kuang, Mingjing Li, Guonian Jin, Zhiyong Wang, and Yang-Cai Xiao. 2025. Implementation of a Student Psychological Crisis Early Warning System Based on Multimodal Data Fusion. *IEEE Transactions on Affective Computing* (2025).

[5] Marcus Kubsch, Sebastian Strauß, Adrian Grimm, Sebastian Gombert, H. Drachsler, Knut Neumann, and N. Rummel. 2025. Self-regulated Learning in the Digitally Enhanced Science Classroom: Toward an Early Warning System. *Journal of the Learning Sciences* (2025).

[6] Ahmed M. Megreya and Ahmed Al-Emadi. 2026. Concurrent and Longitudinal Predictions of Math Anxiety, Science Anxiety, and Enrollment in Science and Art Tracks in Secondary Education. *Educational Psychology* (2026).

[7] Anca O. Muresan, M. Cardei, and I. Cardei. 2025. Exploring Temporal Heterogeneous Graph Deep Learning and Machine Learning Models for Predicting Student Success. *IEEE Transactions on Learning Technologies* (2025).

[8] Behnam Parsaeifard, C. Imhof, Tansu Pancar, I. Comsa, Martin Hlosta, Nicole Bergamin, and P. Bergamin. 2025. Detection of Disengagement from Voluntary Quizzes: An Explainable Machine Learning Approach in Higher Distance Education. *British Journal of Educational Technology* (2025).

[9] Ridley Setiawan, Edi Nursasongko, Abdul Syukur, Fikri Budiman, and Dede Kurniadi. 2025. Handling Class Imbalance in Student Success Prediction Using Machine Learning: A Comparison of SMOTE and SMOTETomek. *IEEE Access* (2025).

[10] J. W. Yeung. 2024. The Dynamic Relationships between Educational Expectations and Science Learning Performance among Students in Secondary School and Their Later Completion of a STEM Degree. *Journal of Educational Psychology* (2024).