Scientific Research Publishing

Voice of the Publisher, 2026, 12(4), 689-701
https://www.scirp.org/journal/vp
ISSN Online: 2380-7598
ISSN Print: 2380-7571

Predicting STEM Persistence and Attrition among U.S. Students: A Longitudinal Machine Learning Analysis of HSLS:09

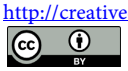
Musa Oladipupo Tijani[1], Lilian A. Oduor[1], Oludotun Oluwatoni Akinsola[2], Gentle Paul Engworo[3], Amen Edoha Engworo[1], Akinkunmi Abdulyeqeen Amoo[4]

[1]Department of Mathematics, Missouri State University, Springfield, MO, USA
[2]Department of Educational Leadership, Research and Technology, Western Michigan University, Kalamazoo, MI, USA
[3]Department of Mathematics, Missouri University of Science and Technology, Rolla, MO, USA
[4]Department of Chemical Engineering, University of Lagos, Lagos, Nigeria
Email: tijanimusatom@gmail.com

How to cite this paper: Tijani, M. O., Oduor, L. A., Akinsola, O. O., Engworo, G. P., Engworo, A. E., & Amoo, A. A. (2026). Predicting STEM Persistence and Attrition among U.S. Students: A Longitudinal Machine Learning Analysis of HSLS:09. *Voice of the Publisher, 12,* 689-701.
https://doi.org/10.4236/vp.2026.124038

Received: September 10, 2026
Accepted: November 1, 2026
Published: November 4, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Open Access

Abstract

Science, technology, engineering, and mathematics (STEM) persistence is critical to the development of the U.S. scientific and technical workforce. However, a significant portion of students who initially intend to pursue STEM fields do not continue into these majors during postsecondary education. The primary objective is to evaluate whether high-school academic achievement, socioeconomic status, demographic characteristics, and motivational measures related to mathematics and science can distinguish students who initially intended to pursue a STEM major and subsequently had a STEM undergraduate major from those whose later undergraduate major was classified as non-STEM. Using data from the High School Longitudinal Study of 2009 (HSLS:09), the study evaluates four predictive approaches: logistic regression, Elastic Net regularized logistic regression, random forest, and Extreme Gradient Boosting (XGBoost). The analytical cohort (N = 2331) revealed an attrition rate of 32.56%. The findings indicate that regularized regression performs competitively with complex tree-based algorithms, achieving a test ROC-AUC of 0.643. Key predictors include mathematics achievement, socioeconomic status, mathematics identity, and science self-efficacy. Results highlight the importance of threshold calibration in early-warning systems, as optimized thresholds significantly improved the detection of students at risk of attrition (sensitivity = 70.4%). The study provides a framework for using longitudinal educational data to support early identification and targeted interventions for at-risk STEM students.

Keywords

STEM Persistence, STEM Attrition, Machine Learning, Predictive Analytics, Elastic Net, XGBoost, HSLS:09

1. Introduction

The development of a STEM workforce is an important concern in the United States. Many students express an intention to pursue STEM during secondary school, but their later postsecondary major may differ from that intention. Research describes movement away from STEM fields as STEM attrition (Chen, 2013; Aulck et al., 2017).

STEM attrition is not a uniform process; it is influenced by a complex interplay of academic preparation, socioeconomic circumstances, and psychological factors. Traditional educational research has frequently relied on descriptive statistics or cross-sectional analyses to identify these factors. However, longitudinal datasets, which track students over multiple years, offer a more powerful opportunity to understand the temporal nature of student trajectories (Duprey et al., 2018). By leveraging data collected early in a student's educational career, researchers can develop predictive models that serve as early-warning systems, allowing for timely interventions before a student leaves the STEM pathway (Aulck et al., 2017; Lee & Chung, 2019).

This study uses the High School Longitudinal Study of 2009 (HSLS:09) to examine the transition from initial high-school STEM-major intention to later undergraduate-major status. The analysis is restricted to the observed eligible cohort; the unweighted results are not treated as nationally representative estimates.

2. Research Statement

The primary objective is to evaluate whether high-school academic achievement, socioeconomic status, demographic characteristics, mathematics and science related motivational measures distinguish students with an initial STEM-major intention whose later undergraduate major was classified as STEM from those whose later major was classified as non-STEM. This outcome represents STEM-intention realization, not uninterrupted enrollment or degree completion. The study addresses the following questions:

1) How well do the evaluated models discriminate between students whose later undergraduate major was classified as STEM and those whose later major was classified as non-STEM?

2) Which academic, socioeconomic, demographic, and motivational predictors contribute most to model predictions?

3) How does selecting a classification threshold using cross-validation affect the detection of students whose later major was classified as non-STEM?

By answering these questions, this study aims to provide a data-driven framework for educational institutions to identify at-risk students and implement targeted support strategies.

3. Literature Review

The literature on STEM persistence and attrition is extensive, spanning decades of research into the factors that support or hinder student success. Researchers have identified several key domains of influence, including academic preparation, demographic disparities, and motivational constructs.

3.1. Academic and Demographic Predictors

Academic preparation, particularly in mathematics and science, is consistently identified as a primary predictor of STEM success. Prior research indicates that high school GPA, standardized test performance, and the rigor of prior mathematics and science preparation are important indicators of whether a student will persist in a STEM major (Chen, 2013, 2015; Chen et al., 2018; Mason et al., 2018; Suárez et al., 2021). For example, early performance in mathematics courses and math proficiency are critical filters; students who enter college with stronger quantitative readiness are more likely to remain in technical fields (Chen, 2013, 2015; Middleton et al., 2019).

Demographic factors also play a significant role in attrition patterns. Studies have repeatedly shown that persistence in STEM varies across gender, race/ethnicity, first-generation status, and socioeconomic circumstances (Andersen & Ward, 2014; Chen, 2013; Suárez et al., 2021). Socioeconomic status (SES) remains an important contextual factor because access to academic preparation, enrichment opportunities, and institutional resources is not evenly distributed across students (Davis & Wilson-Kennedy, 2023; Hansen et al., 2024).

3.2. Motivational and Social Factors

Beyond academic and demographic variables, psychological and motivational factors are essential to understanding the STEM pipeline. Self-efficacy—the belief in one's ability to succeed in specific tasks—is an important component of persistence, while expectancy-value beliefs shape students' plans to continue in STEM (Andersen & Ward, 2014; Wang et al., 2022). Students with a strong sense of STEM or science identity are also more likely to sustain engagement in STEM environments (Lockhart et al., 2022; Seyranian et al., 2018).

Interest and utility value—the perception that STEM is useful and relevant to one's goals—also contribute to persistence (Asher et al., 2023; Hsieh & Simpkins, 2022). Conversely, a weak sense of belonging can undermine persistence even among academically capable students. Recent evidence highlights the importance of belonging, academic hope, identity, and cultural resources for students from groups historically underrepresented in STEM (Davis & Wilson-Kennedy, 2023; Hansen et al., 2024).

3.3. Predictive Modeling in STEM Persistence

Recent advancements in educational data mining and machine learning have shifted the focus toward predictive analytics. Studies have employed logistic regression, neural networks, random forests, and other machine-learning approaches to identify students at risk of STEM attrition or non-retention (Alkhasawneh & Hargraves, 2014; Aulck et al., 2017; Chen et al., 2018; Mason et al., 2018; Sage et al., 2021).

Although predictive models can achieve strong overall classification performance, such results can be misleading when the outcome classes are imbalanced. A model may correctly classify many persisters while failing to identify the smaller group of students who actually leave STEM. Consequently, attrition studies increasingly emphasize class-specific performance and early-alert usefulness rather than overall accuracy alone (Chen et al., 2018; Sage et al., 2021). Longitudinal approaches further show that attrition risk develops over time and can be informed by academic preparation, engagement, and motivational trajectories (Middleton et al., 2019; Sage et al., 2021).

4. Research Method and Procedures

This study follows a systematic pipeline for predictive modeling, utilizing a large-scale longitudinal dataset to train and evaluate multiple machine learning algorithms.

4.1. Data Source and Participants

The data were obtained from the High School Longitudinal Study of 2009 (HSLS:09), conducted by the National Center for Education Statistics (Duprey et al., 2018). HSLS:09 followed a cohort of students from their 9th-grade year through postsecondary education and into the workforce. The original dataset contained 23,503 observations.

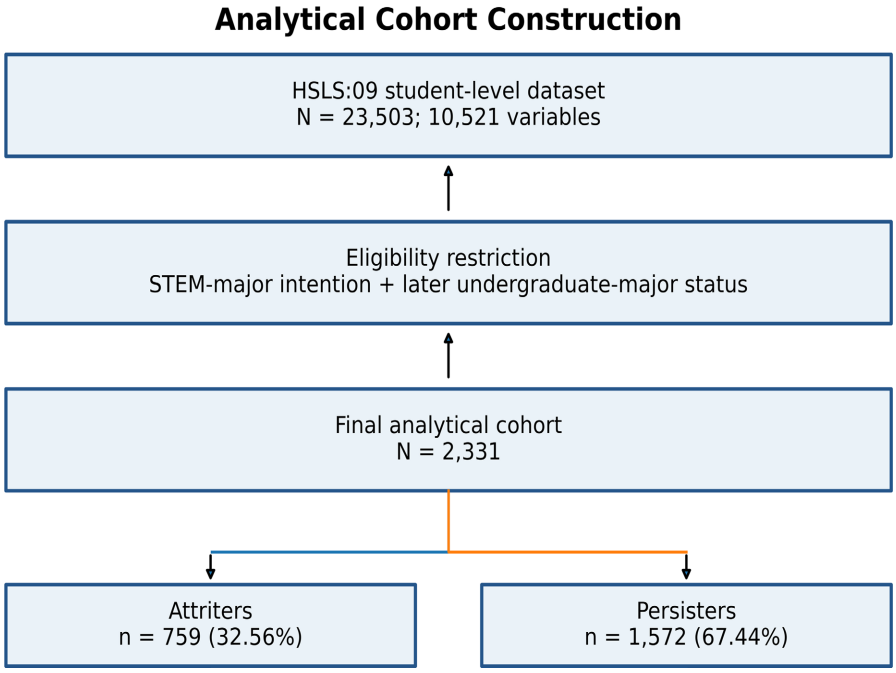
The analytical cohort was restricted to students meeting the initial STEM-intention criterion (`S3FIELD_STEM = 1`) and having valid later undergraduate-major information (`X4RFDGMJSTEM`). Of the 23,503 original records, 2935 met the STEM-intention criterion. After requiring valid later-major information, 2331 students remained; 604 were not retained at this later-major eligibility stage. The final cohort included 1572 students (67.44%) with a later STEM major and 759 (32.56%) with a later non-STEM major. The outcome does not measure continuous STEM enrollment or degree completion. The cohort construction is summarized in **Figure 1**.

1) Indicated an intention to pursue a STEM major during their high school years.

2) Had valid information regarding their later undergraduate major status.

The final sample consisted of 2331 students. These students were classified into two groups:

- Persisters: Students whose later reference-degree major remained in a STEM field.
- Attriters: Students whose later major was in a non-STEM field.



Note. The 604 records were not retained at the valid-later-major eligibility step; the available analysis does not subdivide them into specific missingness categories.

Figure 1. Analytical cohort construction.

4.2. Variables and Measures

The predictor set comprised mathematics achievement (`X1TXMTH`), socioeconomic status (`X1SES`), sex (`X1SEX`), race/ethnicity (`X1RACE`), parent educational expectations (`X1PAREDEXPCT`), mathematics identity (`X1MTHID`), mathematics interest (`X1MTHINT`), mathematics self-efficacy (`X1MTHEFF`), science identity (`X1SCIID`), science interest (`X1SCIINT`), science self-efficacy (`X1SCIEFF`), and science utility (`X1SCIUTI`). Sex, race/ethnicity, and parent educational expectations were treated as categorical predictors; the remaining listed predictors were treated as numeric. GPA was not included in the model predictor set.

1) Academic achievement: standardized mathematics achievement score (`X1TXMTH`).

2) Socioeconomic status: HSLS:09 composite measure (`X1SES`).

3) Demographics: sex (`X1SEX`) and race/ethnicity (`X1RACE`).

4) Motivational measures: mathematics identity, interest, and self-efficacy; science identity, interest, self-efficacy, and utility; and parent educational expectations (`X1PAREDEXPCT`).

Missing numeric predictor values were imputed using medians and categorical values using modes estimated from the training data to avoid data leakage. Preprocessing also included factor handling, dummy encoding, novel-level handling, zero-variance removal, and normalization, with training-derived preprocessing applied to the test data.

4.3. Model Development and Evaluation

Four predictive models were evaluated and compared:

1) Logistic Regression: A standard statistical baseline for binary classification.

2) Elastic Net Regularized Logistic Regression: A model that incorporates both L1 (Lasso) and L2 (Ridge) penalties to handle potential multicollinearity and perform feature selection (Zou & Hastie, 2005).

3) Random Forest: A non-linear ensemble method that builds multiple decision trees to improve stability and predictive power (Breiman, 2001).

4) Extreme Gradient Boosting (XGBoost): An advanced gradient boosting framework known for high performance and efficiency in structured data tasks (Chen & Guestrin, 2016).

The data were divided into stratified training and test samples using an 80:20 split (n = 1864 and n = 467, respectively). Model development used repeated stratified five-fold cross-validation with two repetitions and random seed 123. ROC-AUC was the primary model-selection metric. The models were logistic regression, Elastic Net, random forest, and XGBoost. The reported models did not incorporate NCES sampling weights or complex-survey design adjustments.

- ROC-AUC: Area under the Receiver Operating Characteristic curve.
- PR-AUC: Area under the Precision-Recall curve.
- Sensitivity (Recall): The ability to correctly identify students who will attrite.
- Specificity: The ability to correctly identify students who will persist.
- Precision: The accuracy of attrition predictions.
- F1-Score: The harmonic mean of precision and recall.

For the Elastic Net model, candidate classification thresholds from 0.10 to 0.60 in increments of 0.01 were evaluated using cross-validation predictions. Because repeated cross-validation produced multiple predictions per observation, probabilities were averaged by observation. The threshold maximizing cross-validated balanced accuracy (approximately 0.604) was 0.30. The selected cutoff was then applied to the held-out test sample.

5. Result Analysis

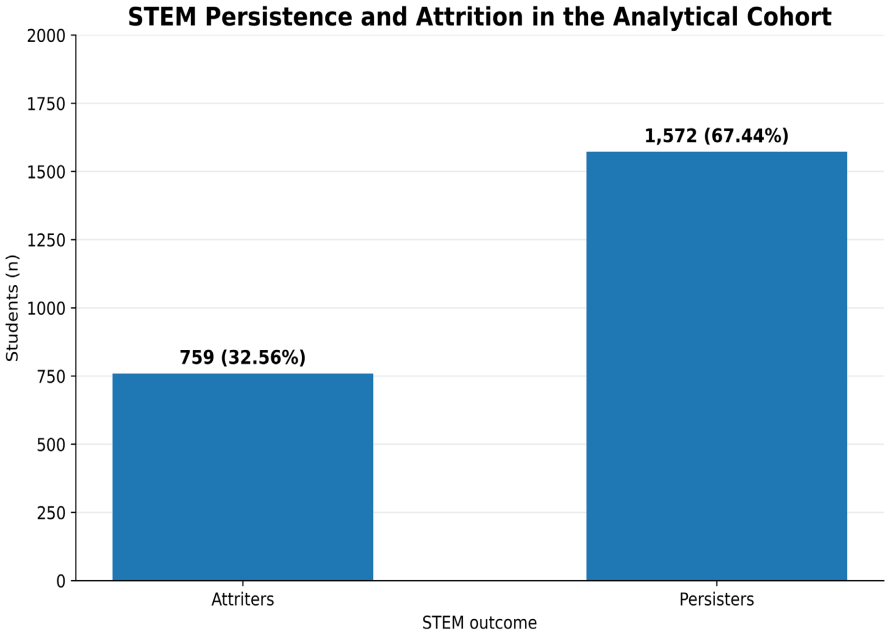
5.1. Descriptive Statistics

The final cohort (N = 2331) included 1572 students (67.44%) with a later STEM major and 759 (32.56%) with a later non-STEM major. The analysis therefore concerns the realization of an initial STEM intention in later major status. The distribution of these two outcome groups is shown in **Figure 2**.

5.2. Model Performance Comparison

The predictive discrimination was moderate. Elastic Net had the highest cross-validation ROC-AUC (0.643; SE = 0.008). The independent test-set comparison is summarized in **Table 1**. Classification metrics at the selected Elastic Net threshold are presented separately in **Table 2**. **Figure 3** provides a visual comparison of the

independent test-set performance across the four models.



Note. The bars show the number and proportion of students with a later non-STEM major (attriters) and a later STEM major (persisters) in the eligible analytical cohort.

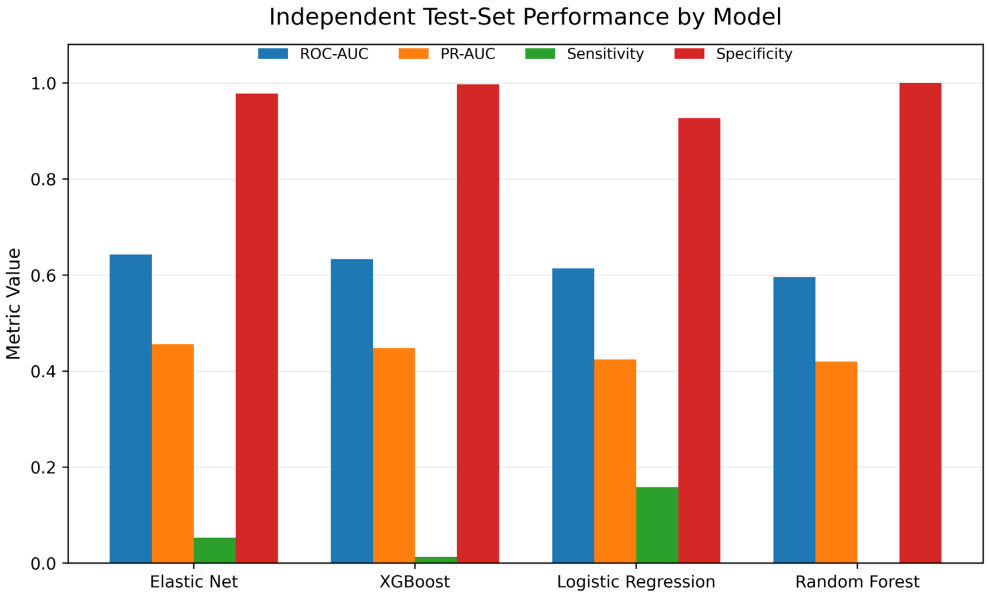
Figure 2. Distribution of STEM-intention outcomes in the analytical cohort.

Table 1. Cross-validation and default-threshold test performance by model.

Model	CV ROC-AUC (SE)	Test ROC-AUC	Test PR-AUC	Test accuracy	Test sensitivity	Test specificity
Elastic Net	0.643 (0.008)	0.643	0.456	0.677	0.053	0.978
XGBoost	0.641 (0.010)	0.633	0.448	0.677	0.013	0.997
Logistic Regression	0.637 (0.008)	0.614	0.424	0.677	0.158	0.927
Random Forest	0.636 (0.008)	0.596	0.420	0.675	0.000	1.000

Table 2. Elastic Net held-out test performance at the selected threshold (0.30).

Metric	Value
Threshold	0.30
ROC-AUC	0.643
PR-AUC	0.456
Accuracy	0.550
Sensitivity (recall)	0.704
Specificity	0.476
Precision	0.393
F1 score	0.505
Balanced accuracy	0.590

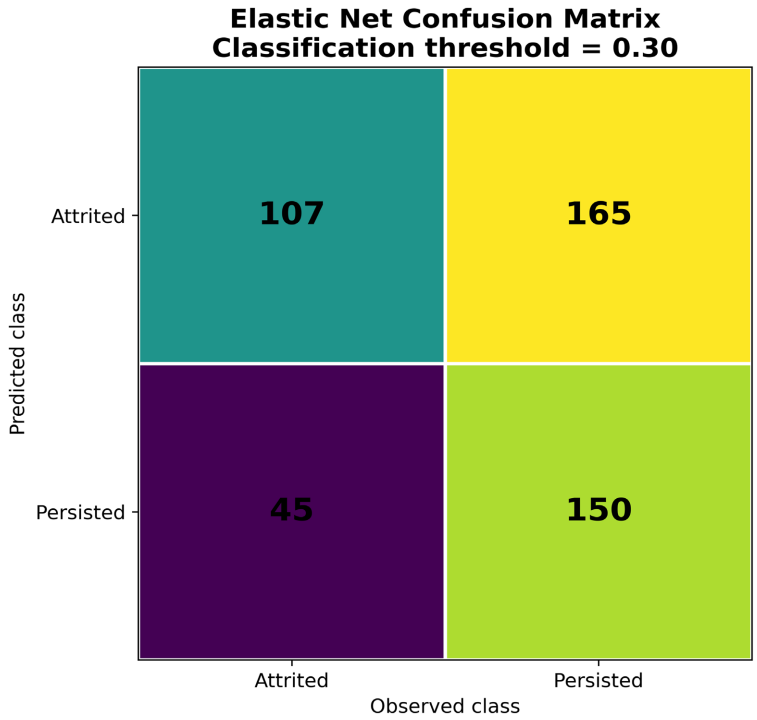


Note. Sensitivity and specificity in this figure are based on the default classification cutoff, not the optimized threshold.

Figure 3. Independent test-set performance by model.

Performance Metrics for the Optimized Elastic Net Model (*Threshold* = 0.3)

The corresponding classification results at the selected threshold are illustrated in Figure 4.



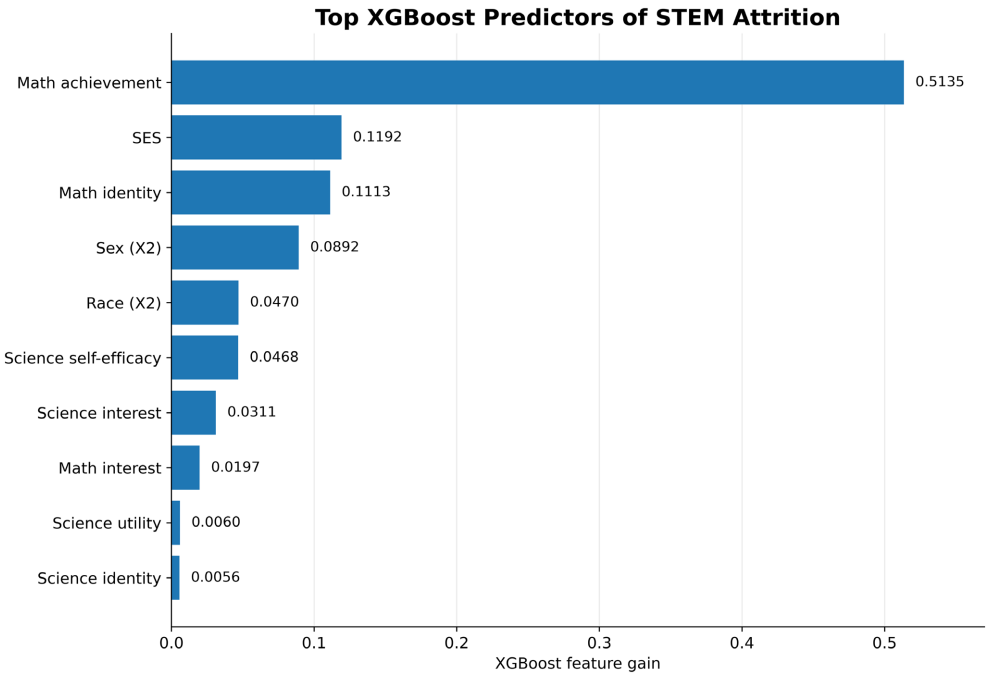
Note. Rows represent predicted class and columns represent observed class; the selected cutoff is 0.30.

Figure 4. Elastic net confusion matrix at threshold 0.30.

Notably, regularized regression (Elastic Net) performed competitively with the more complex tree-based models (Random Forest and XGBoost), suggesting that the predictive signal in this dataset is effectively captured by an additive model.

5.3. Predictor Importance

Feature importance was examined for the tree-based models. The values are model-specific measures of predictive contribution and should not be interpreted as causal effects. Figure 5 presents the XGBoost feature-gain values, while Table 3.



Note. Bars show XGBoost feature gain, a model-specific predictive importance measure, not a causal effect.

Figure 5. Top XGBoost predictors of STEM attrition.

Table 3. Leading feature-importance values from XGBoost and random forest.

Rank	XGBoost predictor	XGBoost gain	Random Forest predictor	Random Forest permutation importance
1	Math Achievement	0.5135	Math Achievement	0.00274
2	Socioeconomic Status	0.1192	Math Identity	0.00186
3	Math Identity	0.1113	Science Self-Efficacy	0.00107
4	Sex_X2	0.0892	Socioeconomic Status	0.00106
5	Race_X2	0.0470	Sex_X2	0.00105
6	Science Self-Efficacy	0.0468	Race_X2	0.00092
7	Science Interest	0.0311	Math Self-Efficacy	0.00069
8	Math Interest	0.0197	Science Interest	Not in reported top entries

compares the leading predictors identified by XGBoost and Random Forest. The XGBoost feature-gain results are visualized in Figure 5.

Top Predictors Ranked by XGBoost and Random Forest

Mathematics achievement emerged as the strongest predictor across all models, emphasizing the foundational role of quantitative skills in STEM persistence (Chen, 2013; Chen, 2015; Larson et al., 2015). Motivational variables, specifically math identity and science self-efficacy, also ranked highly, indicating that how students perceive their own abilities is a critical indicator of their future academic path (Larson et al., 2015; Stewart et al., 2020).

6. Discussion

The results show moderate discrimination for distinguishing students with STEM versus non-STEM later majors in this observed cohort. Elastic Net achieved a test ROC-AUC of 0.643. Because the outcome captures later major status rather than continuous enrollment or completion, it should be interpreted as STEM-intention realization.

The high ranking of mathematics achievement aligns with prior research identifying quantitative preparation and early academic performance as important components of STEM persistence (Chen, 2013, 2015; Chen et al., 2018; Middleton et al., 2019). However, the inclusion of math identity and science self-efficacy among the top predictors suggests that academic performance alone does not tell the full story. Previous studies have similarly demonstrated that expectancy-value beliefs, STEM identity, motivational profiles, and sense of belonging are associated with continued engagement in STEM pathways (Andersen & Ward, 2014; Hansen et al., 2024; Hsieh & Simpkins, 2022; Lockhart et al., 2022; Seyranian et al., 2018).

At the default threshold of 0.50, Elastic Net sensitivity for identifying attriters was low. At the cross-validation-selected threshold of 0.30, test sensitivity was 0.704 and specificity was 0.476. This trade-off should be considered in light of intended use, available support capacity, and the consequences of false-positive and false-negative classifications. The corresponding confusion matrix is reported in Table 4.

Table 4. Elastic Net confusion matrix on the test set (threshold = 0.30).

Predicted class	Actual attrited	Actual persisted
Predicted attrited	107	165
Predicted persisted	45	150

7. Recommendations

Based on the findings of this study, several recommendations are proposed for educational policy and institutional practice:

1) Early Identification: Institutions should use longitudinal data to identify at-

risk students as early as possible. Predictive models, even those with moderate accuracy, can help flag students who might benefit from extra advising or mentoring. (Ogundeji et al., 2022)

2) Holistic Support: Interventions should address both academic and motivational needs. Tutoring programs for mathematics achievement should be paired with efforts to build student confidence and STEM identity.

3) Threshold Calibration: When implementing predictive analytics, administrators must carefully select classification thresholds based on the specific goals of the intervention. A focus on sensitivity is recommended for identifying students for low-cost, high-reach support programs.

4) Integration with Faculty Judgment: Predictive models should serve as decision-support tools for educators and advisors, rather than as deterministic classifiers. Algorithms should supplement, not replace, human expertise (Lee & Chung, 2019).

8. Conclusion

This study evaluated four machine-learning approaches for distinguishing students with an initial STEM intention whose later undergraduate major was classified as STEM from those whose later major was classified as non-STEM. In the observed analytical cohort, Elastic Net produced a test ROC-AUC of 0.643. Mathematics achievement, socioeconomic status, mathematics identity, and science self-efficacy were prominent predictors in the tree-based models. Findings are predictive rather than causal and are not nationally representative estimates because sampling weights and complex-survey adjustments were not incorporated.

The findings may inform the development and evaluation of student-support decision tools, but the model should not be used as a deterministic classifier. Future studies should examine alternative outcome definitions, survey-weighted or design-adjusted methods where feasible, external validation, and performance in real support settings.

Author Contributions

Musa Oladipupo Tijani: Conceptualization, methodology, data curation, formal analysis, machine-learning modeling, visualization, and writing—original draft. Lilian A. Odour: manuscript review, analysis and editing. Oludotun Oluwatoni Akinsola: manuscript review, data curation and editing. Gentle Paul Engworo: manuscript review, visualization and editing. Amen Edoha Engworo: manuscript review, methodology and editing. Amoo Abdulyeqeen Akinkunmi: manuscript review, machine-learning modelling and editing. All authors reviewed the manuscript and approved the final version.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a Hybrid Model to Predict Student First Year Retention in STEM Disciplines Using Machine Learning Techniques. *Journal of STEM Education, 15,* 35-42.

Andersen, L., & Ward, T. J. (2014). Expectancy-Value Models for the STEM Persistence Plans of Ninth-Grade, High-Ability Students: A Comparison between Black, Hispanic, and White Students. *Science Education, 98,* 216-242. https://doi.org/10.1002/sce.21092

Asher, M. W., Harackiewicz, J. M., Beymer, P. N., Hecht, C. A., Lamont, L. B., Else-Quest, N. M. et al. (2023). Utility-Value Intervention Promotes Persistence and Diversity in STEM. *Proceedings of the National Academy of Sciences of the United States of America, 120,* e2300463120. https://doi.org/10.1073/pnas.2300463120

Aulck, L., Aras, R., Li, L., L'Heureux, C., Lu, P., & West, J. D. (2017). *Stem-Ming the Tide: Predicting STEM Attrition Using Student Transcript Data.* arXiv: 1708.09344.

Breiman, L. (2001). Random Forests. *Machine Learning, 45,* 5-32. https://doi.org/10.1023/a:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM. https://doi.org/10.1145/2939672.2939785

Chen, X. (2013). *STEM Attrition: College Students' Paths into and out of STEM Fields (NCES 2014-001).* National Center for Education Statistics.

Chen, X. (2015). STEM Attrition among High-Performing College Students: Scope and Potential Causes. *Journal of Technology and Science Education, 5,* 41-59. https://doi.org/10.3926/jotse.136

Chen, Y., Johri, A., & Rangwala, H. (2018). Running Out of Stem. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 270-279). ACM. https://doi.org/10.1145/3170358.3170410

Davis, R. D., & Wilson-Kennedy, Z. (2023). Leveraging Cultural Wealth, Identities and Motivation: How Diverse Intersectional Groups of Low-Income Undergraduate STEM Students Persist in Collegiate STEM Environments. *Education Sciences, 13,* Article 888. https://doi.org/10.3390/educsci13090888

Duprey, M. A., Pratt, D. J., Jewell, D. M., Cominole, M. B., Fritch, L. B., Ritchie, E. A., Rogers, J. E., Wescott, J. D., & Wilson, D. H. (2018). *High School Longitudinal Study of 2009 (HSLS:09) Base-Year to Second Follow-Up Data File Documentation (NCES 2018-140).* National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Hansen, M. J., Palakal, M. J., & White, L. (2024). The Importance of STEM Sense of Belonging and Academic Hope in Enhancing Persistence for Low-Income, Underrepresented STEM Students. *Journal for STEM Education Research, 7,* 155-180. https://doi.org/10.1007/s41979-023-00096-8

Hsieh, T., & Simpkins, S. D. (2022). The Patterns of Adolescents' Math and Science Motivational Beliefs: Examining Within-Racial/Ethnic Group Changes and Their Relations to STEM Outcomes. *AERA Open, 8.* https://doi.org/10.1177/23328584221083673

Larson, L. M., Pesch, K. M., Surapaneni, S., Bonitz, V. S., Wu, T., & Werbel, J. D. (2015). Predicting Graduation: The Role of Mathematics/Science Self-Efficacy. *Journal of Career Assessment, 23,* 399-409. https://doi.org/10.1177/1069072714547322

Lee, S., & Chung, J. Y. (2019). The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. *Applied Sciences, 9,* Article 3093. https://doi.org/10.3390/app9153093

Lockhart, M. E., Kwok, O., Yoon, M., & Wong, R. (2022). An Important Component to Investigating STEM Persistence: The Development and Validation of the Science Identity (SCIID) Scale. *International Journal of STEM Education, 9,* Article No. 34. https://doi.org/10.1186/s40594-022-00351-1

Mason, C., Twomey, J., Wright, D., & Whitman, L. (2018). Predicting Engineering Student Attrition Risk Using a Probabilistic Neural Network and Comparing Results with a Backpropagation Neural Network and Logistic Regression. *Research in Higher Education, 59,* 382-400. https://doi.org/10.1007/s11162-017-9473-z

Middleton, J. A., Mangu, D., & Lee, A. (2019). A Longitudinal Study of Mathematics and Science Motivation Patterns for Stem-Intending High Schoolers in the Us. In M. Hannula, G. Leder, F. Morselli, M. Vollstedt, & Q. Zhang (Eds.), *Affect and Mathematics Education* (pp. 89-105). Springer. https://doi.org/10.1007/978-3-030-13761-8_5

Ogundeji, R. K., Akinyemi, J. A., & Tijani, M. (2022). Naive Bayes Algorithm for Document Classification. *Annals of Mathematics and Computer Science, 7,* 54-65.

Sage, A. J., Cervato, C., Genschel, U., & Ogilvie, C. A. (2021). Combining Academics and Social Engagement: A Major-Specific Early Alert Method to Counter Student Attrition in Science, Technology, Engineering, and Mathematics. *Journal of College Student Retention: Research, Theory & Practice, 22,* 611-626. https://doi.org/10.1177/1521025118780502

Seyranian, V., Madva, A., Duong, N., Abramzon, N., Tibbetts, Y., & Harackiewicz, J. M. (2018). The Longitudinal Effects of STEM Identity and Gender on Flourishing and Achievement in College Physics. *International Journal of STEM Education, 5,* Article No. 40. https://doi.org/10.1186/s40594-018-0137-0

Stewart, J., Henderson, R., Michaluk, L., Deshler, J., Fuller, E., & Rambo-Hernandez, K. (2020). Using the Social Cognitive Theory Framework to Chart Gender Differences in the Developmental Trajectory of STEM Self-Efficacy in Science and Engineering Students. *Journal of Science Education and Technology, 29,* 758-773. https://doi.org/10.1007/s10956-020-09853-5

Suárez, M. I., Dabney, A. R., Waxman, H. C., Scott, T. P., & Bentz, A. O. (2021). Exploring Factors That Predict STEM Persistence at a Large, Public Research University. *International Journal of Higher Education, 10,* Article 161. https://doi.org/10.5430/ijhe.v10n4p161

Wang, X., Dourado, A., Thomas, P. B., & Bego, C. R. (2022). Modeling Engineering Persistence through Expectancy Value Theory and Machine Learning Techniques. In *2022 IEEE Frontiers in Education Conference (FIE)* (pp. 1-9). IEEE. https://doi.org/10.1109/fie56618.2022.9962581

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 67,* 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x