The Future of Engineering Education
2024 Annual Conference & Exposition
Oregon Convention Center
Portland, OR . June 23 - 26, 2024
ASEE
Paper ID #41785

Integrating Data-Driven and Career Development Theory-Driven Approaches to Study High School Student Persistence in STEM Career Aspirations

tonghui xu, University of Massachusetts, Lowell

PhD student

Dr. Hsien-Yuan Hsu, University of Massachusetts, Lowell

Dr. Hsien-Yuan Hsu is an Assistant Professor in Research and Evaluation in the College of Education at the University of Massachusetts Lowell. Dr. Hsu received his PhD in Educational Psychology from Texas A&M University and has a background of statistics

Integrating Features Selection and Career Development Theory-Driven Approaches to Study High School Student Persistence in STEM Career Aspirations

Abstract

Educational researchers often rely on the theory-driven approach to identify predictors from large-scale survey (LSS) data. Applying a single theory may potentially overlook valuable predictors. Using multiple theories for predictor identification can result in excessive predictors, complicating model specification and parameter estimation. Feature selection is a common machine learning algorithm used to identify important predictors for data analysis, but sometimes the feature selection results may provide uninterpretable results. Therefore, this study employed a blended approach that integrated data-driven and multiple-theory-driven approaches. This blended approach can help researchers reduce predictors and retain interpretability. Based on that, we conducted a study to analyze the Education Longitudinal Study of 2002 (ELS:2002) dataset to establish a procedure for identifying predictors. Multilevel modeling was then utilized to study high school students' persistence in STEM career aspirations.

KEYWORDS:

feature selection, STEM career aspirations, career theory, large scale survey, multilevel molding

Introduction

High school students' aspirations for STEM occupations can significantly influence their decisions to pursue a STEM track in college or as a career. In 2016, there were only 568,000 STEM graduates in the U.S., compared to 2.6 million in India and 4.7 million in China [1]. STEM literacy is critical to human capital competency for the economy [2]. Therefore, encouraging more high school students to aspire to STEM careers can increase the likelihood of applying for jobs in STEM fields. Because many internal and external factors may influence high school students' aspirations for STEM careers, previous research on this topic often employs a theory-driven approach to identify predictors from large scale survey (LSS) data and formulate hypotheses for statistical tests. Existing LSS datasets, such as the Education Longitudinal Study of 2002 (ELS:2002), promise a comprehensive investigation of the factors that contribute to high school students' persistence in STEM career aspirations. According to our literature review, the career theories explaining persistence in STEM career aspirations include social cognitive career theory (SCCT) [3], expectancy-value theory (EVT) [4], and expectation states theory (EST) [5].

Identifying variables based on the theory is a commonly used practice because it can increase the opportunity to explain the phenomena of interest, rather than simply describe them [6].

Many studies suggested that these theories can be applied to the study of high school student career aspirations. For example, the EVT was used to select predictors from the Programme for International Student Assessment 2012 dataset to study gender differences in the rate of student' aspirations to STEM occupations [7], the SCCT can be used to select predictors from the High School Longitudinal Study of 2009-2014 in a candidate variable subset [8], and the EST can be applied to independent variables from the NCES Education Longitudinal Study 1988-2000 dataset to study the gender differences and determined the role of the high school context in STEM majors' plans [9]. However, when using the theory-driven approach with large-scale dataset, challenges emerge. Many studies tend to rely on one theory to identify predictors, potentially missing out on the rich insights these datasets offer. Yet, employing multiple theories for predictor identification can lead to an overwhelming number of predictors. This is where the data-driven approach becomes beneficial. We can reduce the number of predictors identified from multiple theories based on the feature selection model. Notably, the predictors selected using this data-driven method remain interpretable since they are originally sourced from established theories.

This study proposes a blended approach that integrates theory-driven and data-driven methods. We demonstrate this approach by analyzing the ELS:2002 dataset to construct a model explaining high school students' persistence in STEM career aspirations. Initially, we use three theories, namely SCCT, EVT, and EST, to identify candidate predictors from ELS:2002 so that we can maximize data utilization. By using the Boruta algorithm, a data-driven method based on random forest classification, we streamline predictor selection from this extensive list to construct the final model.

Feature selection

Feature selection (FS) is a vast and fruitful research field in pattern recognition, machine learning, statistics, and data mining [10]. The benefits of applying feature selection in to select the appropriate feature subset to construct the data science patterns, reduce the running time, understand the relationship between predictive variables and outcomes, and reduction in case of high dimensional datasets [11]. Many studies suggested the FS is an appropriate method to identify the important predictors for data analysis in education fields. For example, integrated

wrapper feature selection method and classification data mining models to identify the important variables to predict the students' performance [12]. Random forest algorithm to analyze the High School Longitudinal Study of 2009 data to identify the important variables which impact the engineering major choice [13].

The Boruta algorithm is a high-performance FS that employs a novel feature selection algorithm based on the random forest (RF) classification learning method. It is available as an R package [14]. RF combines the predictions of multiple individual models to predict outcomes. It is better than the outcomes computed by a single learning model. Typically, RF constructs multiple decision tree models, and each tree runs a random subset of the features in the training dataset independently. In the final step, the RF aggregates the predicted results of all the individual trees as the final outcomes. First, the Boruta algorithm duplicates all records and places them into a new data frame. Within this new frame, the system shuffles the values of each variable using a random permutation method and renames all variables as shadow attributes. Boruta creates shadow features, which are copies of the original variables with their values randomly shuffled. These shadow variables serve to determine the importance of the original variables. Next, the algorithm trains the model using both the original variables and their corresponding shadow variables. In the second step, the RF classifier computes the Z-values of loss accuracy for both shadow and real variables. The algorithm labels a hit for any real attribute with Z-scores better than the maximum Z-value across all shadow attributes (MZSA). Subsequently, the algorithm deletes all shadow attributes and proceeds with another iteration. In each iteration, the same steps are repeated until assigning the final importance score for the attributes or until the algorithm is stopped. At the end of the iterations, Boruta categorizes each feature as: "confirmed" (important), "tentative" (meaning its importance was not significantly better than shadow features), or "rejected" (unimportant). The following section will demonstrate the study of feature selection.

Demonstration

Data Sources

The (ELS:2002) data is an integrated survey and assessment involving multiple respondents sponsored by the National Center for Education Statistics (NCES) [15]. This dataset inflects a descriptive portrait of these tenth-grader high school students from 2002 to 2012. It includes four data collection waves: (1) based year (2002), (2) first follow-up (2004), (3) second

follow-up (2006), (4), and third follow-up (2012). In the first two data collection waves, the based year comprises 10th grades and sophomores in the spring term of the 2001–02 school year and the first follow-up years included 12-grade students in 2004. The second follow-up years collect the data from all respondents had graduate high schools for 2 years and in the third follow-up years, all the respondents had already graduated high schools for 8 years. This dataset is to track the students' academic and career trajectories in different time points.

Sample

In this study, we retained data from the base year and the first follow-up year, revealing that out of 2,741 high school students surveyed in their 10th grade, 56.18% still expressed an interest in STEM careers during their 12th-grade year. Conversely, 43.82% of students shifted their career preferences to non-STEM fields or indicated uncertainty. The analytical sample included 2,741 12th-grade individuals from 360 high schools. Our aim was to identify predictive variables influencing the choice of STEM occupations by respondents at age 30, based on their responses during the 12th-grade year. The sample comprised tenth-grade students, with 42.06% male and 57.94% female, including 10.91% Asian and Hawaiian, 12.08% Black or African American, 10.80% Hispanic, 61.44% White, and 5.87% other ethnicities. Additionally, 23.97% attended private or Catholic schools, while 76.21% were enrolled in public schools.

Variables

This study included two follow-up data: (1) BY variables were the base year data (10th-grade), (2) F1 variables were the first follow-up data (12th-grade). Based on the selected predictive variables from theory-driven approaches (i.e., SCCT, EVT, and SCT) papers, we created 21 predictive variable clusters to identify the related ELS:2002 predictive variables. A total of 81 related ELS:2002 variables were selected from the dataset. Among of these predictive variables, 26 (32.10%) variables are level 2 predictors, and 55 (68.90%) are level 1 predictors.

After that, we distributed these related ELS:2002 variables to the related theory-driven approach variable clusters. Table 1 indicated the selected variables by feature selection methods (Boruta), EVT, EST, and SCCT, along with the variable descriptions from ELS 2002. These variable clusters were based on the related studies (see, e.g., [7], [8], [9], [16], [17], [18], and [19]). For example, in the first row, the variable cluster named "self-efficacy" was selected by EVT and SCCT, and the regression model indicated that self-efficacy is significant. The ELS 2002 data includes two self-efficacy variables: English self-efficacy and math self-efficacy. The

last column counts the total significant variables from the three theories and important variables from the feature selection method.

Table 1

Result of theory driven approaches and Boruta feature selection

Variable clusters	EVT	EST	SCCT	Boruta	ELS:2002 variables
self-efficacy	X*		X*	X*	English and math self-efficacy
Student spend time to study	X			X*	Hours per week spent on homework in school and out of school
Gender	X*	X*	X*	X*	Student gender
Generation status	X				Generational status
High school program			X*		Courses offered by school
Math and science extracurricular	X				Participated in science/math fair and voc/tech skills competition
Math identity			X*		Can learn to be good at math and become totally absorbed in mathematics.
Math performance	X*	X*	X*	X*	F1 math standardized score
Math teacher's positive influence			X*	X*	Teacher's expectations for student education level and support for student success
Math utility and interest	X*	X*	X		Mathematics is important and interesting
Number of siblings			X		Number of siblings 10th-grader has
Parental expectations			X*	X*	Parents' educational expectations and aspirations for their child's future after high school

Parent's education level			X		Mother and father's highest level of education-composite
Parents have a STEM occupation	X*	X			STEM occupation for mother/female and father/male guardian
Parental involvement			X		Parent participant in the school activities
Student success expectations	X*		X*	X*	Student desires to succeed academically
Student involvement in academic			X	X*	Student effort in studying, belief in the importance of education, and persistence.
Race	X*	X*	X*		Student races
Reading performance		X		X*	Reading test standardized score
School belonging and engagement			X*		Student feeling of the school and safety
School setting			X		School offers facility for students
School type		X	X*		Region and school type
SES	X*	X*	X*	X*	Socio-economic status
Student educational expectations		X*	X*	X*	Expected education level for job requirements and personal educational aspirations.
Total count	8	6	13	11	81 ELS 2002 variables

Note that X* indicates that the theory-driven model selected and marked this variable as significant, while Boruta identified this variable as important. X label meant the theory drive model selected, but not significant.

The dependent variable of this study was to describe the 12-grade students who expect to choose STEM occupations at age 30. The original variable included 9 STEM occupations, not STEM occupations, and the option of I don't know. Therefore, we combined the STEM jobs categories and recreated this variable to become (a) 1: STEM occupation (56.18%) and (b) 0:

non-STEM occupation (43.82%). The non-STEM occupation includes non-STEM jobs, and I don't know.

Data Analysis

We then used the Boruta package in R 4.13 version to analyze the predictive variables selected from the 21 variables clusters, retained the significant variables as candidate predictive variables, and also create boxplot figure to visualize the results of feature selection. Based on the results of Boruta feature selection and theory-driven approaches, we applied multilevel modeling to investigate the relationship between student STEM career persistence and selected important predictors.

To understand the extent to which individual outcomes are influenced by group-level factors, we first computed the intraclass correlation coefficient (ICC) of the unconditional model to assess the total variance across schools and individuals. Next, we employed multilevel modeling to create three distinct random effect models: (1) Molde1: predictors selected by at three theory-driven approaches and Boruta, (2) Molde2: at least two theory-driven approaches and Boruta, and (2) Molde3: at least one theory-driven approach and Boruta.

Results

Boruta Feature Selection

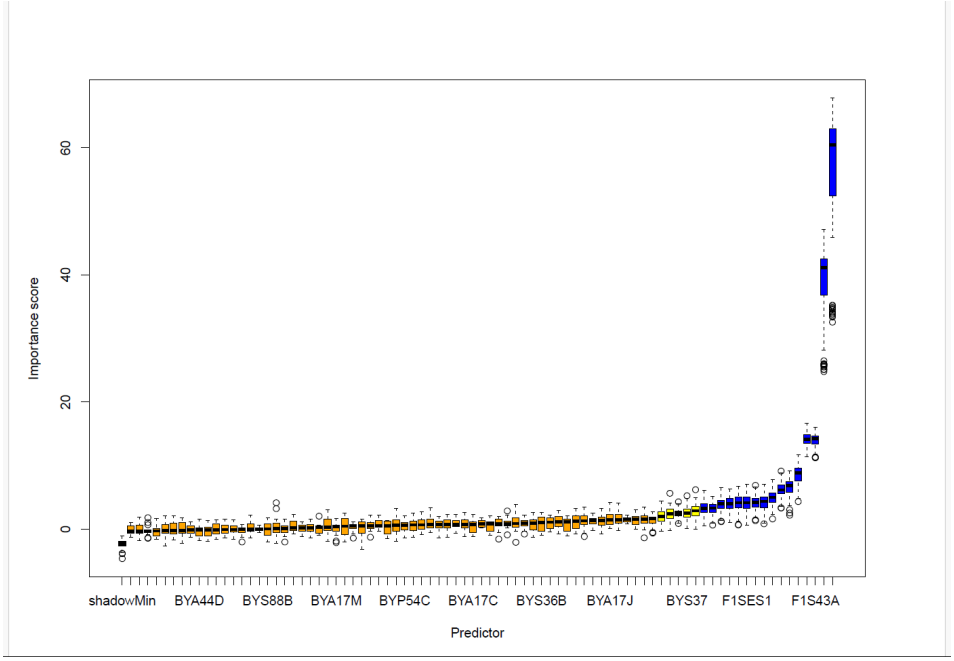
A total of 81 candidate predictive variables were submitted to the Boruta function. The Boruta package can label the final decision of feature selection. Sixteen predictive variables were confirmed as important variables from 11 variable clusters, with 4 labeled as tentative variables, and 61 rejected by the Boruta method. In addition, The Boruta package can compute the importance scores for each variable. The higher value of importance score indicated the variables are more important to the dependent variable. Out of these 17 variables, we selected variables that were suggested by at least one theory-driven approach. Therefore, the total number of selected variables is 12.

Figure 1 showed the boxplot of the results of Boruta. The x-axis was the variable names. Because we had 81 variables, so the *plot* function only presented a few variable names. The y-axis indicated the importance score of each variable. In this figure, we changed the default color setting of the *plot* function. From right to left, the blue boxplot corresponded to predictive variables that were confirmed as uncertain variables, and the orange boxplot corresponds to the

predictive variables that were marked as irrelevant variables. The black boxplot showed the range of minimum value of shadow variable.

Figure 1

Boxplot of Importance Score of Predictors



The 12 variables included (a) English and math self-efficacy, (b) math performance, (c) gender, (d) how in school far mother and father wants student to go, (e) mother's desires for student after high school, (f) how often student discussed jobs with parents, (g) student academic success expectations, (h) social economic status, (i) how far in school student think will get, and (j) how much education student think will be need for job at age 30.

Multilevel Modeling

Table 2 shows the model comparison for multilevel modeling. As the ICC is small, we focused solely on the random intercept model. In comparison with the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and log-likelihood, Model 2 exhibited a lower BIC value. Since BIC favors simplicity and larger sample sizes, we selected Model 2 as the most appropriate model.

Table 2

Model Comparison

	Unconditional Model	Model1	Model2	Model3

AIC	3760.8	3718.9	3397.2	3372.0
BIC	3772.6	3748.4	3456.4	3466.7
Log-likelihood	-1878.4	-1854.4	-1688.6	-1670.0
Num of predictors	0	3	8	14
ICC: 0.01				

Table 3 indicated the statistical results of predictors in Model 2, including variable names, descriptions, odds ratios, standard errors, and importance scores. For instance, in the first row, the variable name is female, the reference group is male, the OR of female is 1.63, standard error is 0.08. and the p-value is less than 0.01. Gender and two education expectation variables were the categorical variables, but we considered education expectation variables as continuous in Model 2.

Table 3

Results of multilevel modeling

Variable name	Descriptions	OR	SE
Female	Gender (reference group: male)	1.63***	0.08
How far in school student thinks will get degree	How far in school respondent thinks he/she will get degree (e.g., high school, graduate from college).	1.52***	0.05
How much education respondent thinks needed for job at age 30	Respondents' perceptions of the level of education required to obtain the job they expect or plan to have by the age of 30.	1.81***	0.04
Student success expectations	Respondents' success expectations in school courses.	1.19**	0.06
Mathematics self-efficacy	Respondent's self-efficacy in math.	1.81***	0.05
English self-efficacy	Respondent's self-efficacy in English.	0.90	0.05
Socio-economic status	Socio-economic status	0.94	0.05
Math performance	Student math performance	1.05	0.04

P-value: 0.05**; 0.01***

Discussion

In the discussion, we identified 12 predictive variables confirmed as important factors by the Boruta method. Among these, 4 predictors were found to be significant across all three theories. Additionally, 4 predictors were significant in two theories, while 6 predictors were found to be significant in one theory. The matching rate between the Boruta method and the three theories is 87.5%. When we ran Model 2, 6 out of 8 predictors were found to be significant. These results suggested that (a) theory-driven approaches (i.e., EVT, SCCT, and EST) can strengthen data-driven approaches (Boruta), and vice versa, when analyzing high-dimensional survey data, and (b) combining data-driven and theory-driven approaches can be effective in identifying important predictors. Based on the multilevel modeling results, we found that 5 predictors are significant. Compared with males, female high school students exhibit higher STEM career persistence, controlling for other variables. Additionally, 12th-grade students with higher education expectations or high mathematics self-efficacy also demonstrate higher STEM career persistence. Furthermore, 10th-grade students with high academic success expectations show higher persistence in STEM careers.

Based on our findings, we proposed instructing students on the significance of perseverance, resilience, and continuous learning. Encourage them to perceive setbacks and challenges as opportunities for personal growth rather than obstacles to success. Furthermore, we advocate for disseminating information regarding the benefits of pursuing post-secondary education, particularly in STEM fields, which can lead to expanded career prospects and higher earning potential. To broaden students' exposure to various STEM careers and pathways, we recommend organizing extracurricular activities, inviting guest speakers, and arranging industry or college visits. Moreover, providing supplementary resources such as tutoring, peer mentoring, and engaging learning activities can enhance students' confidence in their learning, particularly in mathematical skills and abilities. Finally, establishing mentorship programs, workshops, and networking opportunities tailored specifically to address the unique obstacles encountered by male students in persisting with STEM careers can further support their career aspirations.

Conclusion

This study demonstrated the integration of theory-driven approach and feature selection algorithm to study high school student STEM career aspiration at age 30. The results suggested that (1) the integration of the theory-driven approach with a feature selection algorithm is an

effective method for selecting important variables in educational big data studies, particularly when dealing with high-dimensional data., and (2) these results can be interpreted to develop strategies to improve high school students' STEM career aspirations and persistence. We hope this study can inspire more educational researchers to use machine learning algorithms to analyze big educational datasets.

References

[1] McCarthy, N. (2017). Recent graduates in STEMM. https://www.industryweek.com/talent/article/21998889/the-countries-with-the-most-stem-graduates Retrieved 23 April 2021.

[2] Capraro, R. M., & Han, S. (2014). STEM: The education frontier to meet 21st century challenges. *Middle Grades Research Journal*, *9*(3), XV.

[3] Lent, R. W., Brown, S. D., & Hackett, G. (1994). "Toward a unifying social cognitive theory of career and academic interest, choice, and performance," Journal of vocational behavior, 45(1), 79-122.

[4] Eccles, J. S., & Wigfield, A. (2002). "Motivational beliefs, values, and goals," Annual review of psychology, 53(1), 109-132.

[5] Ridgeway, C. L., & Correll, S. J. (2006). "Consensus and the creation of status beliefs," Social Forces, 85(1), 431-453.

[6] Qiu, D., Li, X., Xue, Y., Fu, K., Zhang, W., Shao, T., & Fu, Y. (2023). Analysis and prediction of rockburst intensity using improved DS evidence theory based on multiple machine learning algorithms. *Tunnelling and Underground Space Technology*, *140*, 105331.

[7] Mann, A., & DiPrete, T. A. (2016). "The consequences of the national math and science performance environment for gender differences in STEM aspiration," Sociological Science, 3, 568.

[8] Mau, W. C. J., & Li, J. (2018). "Factors influencing STEM career aspirations of underrepresented high school students," The Career Development Quarterly, 66(3), 246-258.

[9] Legewie, J., & DiPrete, T. A. (2014). The high school environment and the gender gap in science and engineering. *Sociology of education*, *87*(4), 259-280.

[10] Dhanalakshmi, R., & Khaire, U. M. (2019). "Feature selection and classification of microarray data for cancer prediction using mapreduce implementation of random forest algorithm," Journal of Scientific & Industrial Research.

[11] Khaire, U. M., & Dhanalakshmi, R. (2019). Stability of feature selection algorithm: A review. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2019.06.012

[12] Chaudhury, P., & Tripathy, H. K. (2017). "An empirical study on attribute selection of student performance prediction model," International Journal of Learning Technology, 12(3), 241-252.

[13] Tan, L., Main, J. B., & Darolia, R. (2021). "Using random forest analysis to identify student demographic and high school-level factors that predict college engineering major choice," Journal of Engineering Education, 110(3), 572-593.

[14] Kursa, M. B., & Rudnicki, W. R. (2010). "Feature selection with the Boruta package," Journal of statistical software, 36, 1-13.

[15] Ingels, S. J., Pratt, D. J., Wilson, D., Burns, L. J., Currivan, D., Rogers, J. E., & Hubbard-Bednasz, S. (2007). "Education Longitudinal Study of 2002 (ELS: 2002): Base-Year to Second Follow-Up Data File Documentation. NCES 2008-347," National Center for Education Statistics.

[16] Rachmatullah, A., Reichsman, F., Lord, T., Dorsey, C., Mott, B., Lester, J., & Wiebe, E. (2021). "Modeling secondary students' genetics learning in a game-based environment: Integrating the expectancy-value theory of achievement motivation and flow theory," Journal of Science Education and Technology, 30, 511-528.

[17] Mann, A., & DiPrete, T. A. (2016). The consequences of the national math and science performance environment for gender differences in STEM aspiration. *Sociological Science*, *3*, 568.

[18] Mau, W. C., & Bikos, L. H. (2000). Educational and vocational aspirations of minority and female students: A longitudinal study. *Journal of counseling & development*, *78*(2), 186-194.

[19] Andersen, L., & Ward, T. J. (2013). An expectancy-value model for the STEM persistence of ninth-grade, underrepresented minority students. In Community colleges and STEM (pp. 59-74). Routledge.