2023-10-10

Using Machine Learning Methods To Develop Person-Centered Models Predicting STEM Major Choice

Marcell NAGY
Budapest University of Technology and Economics, marcell.nagy94@gmail.com

Joyce MAIN
Purdue University, United States of America, jmain@purdue.edu

Roland MOLONTAY
Budapest University of Technology and Economics, molontay@math.bme.hu

See next page for additional authors

Using Machine Learning Methods to Develop Person-Centered Models Predicting STEM Major Choice

M. Nagy
Department of Stochastics, Institute of Mathematics,
Budapest University of Technology and Economics,
Műegyetem rkp. 3., H-1111 Budapest, Hungary
ORCID: 0000-0001-5666-7777

J. B. Main[1]
School of Engineering Education, Purdue University,
West Lafayette, IN, USA
ORCID: 0000-0002-3984-533X

R. Molontay
Department of Stochastics, Institute of Mathematics,
Budapest University of Technology and Economics,
Műegyetem rkp. 3., H-1111 Budapest, Hungary
ORCID: 0000-0002-0666-5279

A. L. Griffith
Department of Economics, Wake Forest University,
Winston Salem, North Carolina, USA
ORCID: 0000-0003-2538-0460

***Conference Key Area**: Recruitment and Retention of Engineering Students*
***Keywords**: machine learning, STEM major, person-centered models, explainable artificial intelligence (XAI), educational data science*

[1] *Corresponding Author: J. B. Main, jmain@purdue.edu*

ABSTRACT

Understanding the factors that influence the choice of a STEM major is important for developing effective strategies to increase participation in STEM fields and meet the growing demand for skilled workers. This research is based on the nationally representative data of 25,206 students surveyed in the High School Longitudinal Study of 2009 (HSLS:09). The HSLS:09 includes longitudinal data from 9th-grade students through their postsecondary study. First, we use machine learning to predict who is going to opt for a STEM major. Then we use interpretable ML tools, such as SHAP values, to investigate the key factors that influence students' decisions to pursue a college STEM major. We identified with a relatively high degree of accuracy the students who will later choose a STEM major, namely our CatBoost classifier achieved an AUC score of 0.791. Moreover, by interpreting the model, we find that having a science or math identity, as well as demographic characteristics, such as gender and race, play important roles in the decision to pursue a STEM major. For example, Asians are more, females are less likely to consider a STEM major, on the other hand, we also find that gender and race do not influence students' science or math identity.

1 INTRODUCTION

Science, Technology, Engineering, and Mathematics (STEM) fields are critical for innovation, economic growth, and national competitiveness. However, the limited number of students in STEM majors and professions and the underrepresentation of students in these fields is a persistent challenge. To address this, it's essential to understand the factors that influence students' decisions to pursue a STEM major. By identifying these factors, policymakers and educators can develop programs and strategies to increase participation and diversity in STEM fields, meeting the demand for skilled STEM professionals from the workforce.

Several studies have investigated the factors that influence students' decisions to pursue a STEM major. For example, Wang (2013) found that intent to major in STEM is directly affected by 12th-grade math achievement, exposure to math and science courses, and math self-efficacy beliefs. Sahin et al. (2018) found that males and Asian students are more likely to pursue a STEM major. Moreover, they reported that students, who engage in more STEM project-based learning activities, achieve higher GPAs, receive increased encouragement from parents and teachers, exhibit greater math/science efficacy and interest, are more likely to choose STEM majors in college. In a very recent and closely related work by Chang et al. (2023) utilized the HSLS:09 dataset and employed a decision tree to predict STEM major choice. They found that calculus credits, science identity, total STEM credits, and math achievement are the most influential factors during high school years of college STEM major selection. Similarly, Kurban et al. (2019) used structural equation modeling to understand STEM readiness and intention to pursue STEM fields, also by relying on the HSLS dataset. The authors found that STEM major selection is primarily influenced by STEM readiness, math/science interest, and self-efficacy.

Here, we aim to use machine learning (ML) models to predict which students are likely to opt for a STEM major and investigate the key factors that influence students' decisions. To achieve this, similarly to Chang et al. (2023), we analyze the nationally representative HSLS data set, which tracks a cohort of students from the beginning of high school to post-secondary education. By leveraging this data set, we can develop a predictive model that identifies the most critical predictors of STEM major selection.

To gain further insights into the mechanisms underlying our predictive model, we will use interpretable ML/explainable AI tools, such as SHAP values. These tools allow us to identify the most important predictors and how they influence the model's output, i.e., students' decision to pursue a college STEM major.

Previous studies in the field have predominantly relied on classical statistical methods like structural equation modeling, logistic regression, or basic ML techniques such as decision trees. In contrast, here we employ advanced ML techniques, specifically CatBoost for modeling purposes and SHAP values for interpretation, thereby providing a more comprehensive and nuanced analysis of the data.

2 DATA

This study is based on the US nationally representative data of the High School Longitudinal Study of 2009 (HSLS:09). The HSLS:09 includes longitudinal data from 9th-grade students through their postsecondary study. The data were collected in five waves: base year (9th grade), first follow-up (11th grade), high school transcript (12th grade), second follow-up (3 years after high school), and post-secondary transcript (4 years after high school). The variables include the results of surveys (with students, parents, teachers, administrators, and counselors), assessment tests, and transcripts.

The original dataset contains 25,210 rows and 4,014 features, however, there is a great deal of redundancy in the features (e.g., the same questions are asked in multiple collection waves). Hence, to avoid overfitting and to get easily interpretable results we selected a subset of 104 features, aiming to have variables from all groups of variables and to have a relevant but rich set of variables. The selection contains 6 **personal** features (e.g., sex, race, socio-economic status), 8 **high-school related** variables (e.g., geographic region, avg. caseload for counselors), 12 **general** features regarding the students' **personality/expectations/lifestyle** (e.g., the scale of school motivation, the highest level of education student indicated will meet minimum requirements, hours spent playing video games on a typical schoolday), 67 **math and science** related features (e.g., the scale of student's mathematics/science identity, math assessment score, teacher makes science interesting), 10 **transcript** variables (GPA in different courses), and finally a **target variable** that indicates whether the considered major upon postsecondary entry is in a STEM field.

3 METHODOLOGY

3.1 Modeling

In this study, we utilize gradient-boosted tree algorithms, such as XGBoost and CatBoost. These algorithms have been shown to achieve state-of-the-art performance

on tabular datasets as they often outperform the most recent deep learning models (Grinsztajn et al. 2022). Gradient boosting is a type of ensemble learning method that involves combining several decision trees to create a stronger, more accurate model. Here, we assume the reader is familiar with the basic concepts of machine learning, for a great overview see the book of Hastie et al. (2009).

3.2 Evaluation

To evaluate the performance of our models, we employ a 5-fold cross-validation strategy, which involves dividing the dataset into five equal parts and using four parts for training and the remaining part for testing. We repeat this process five times, each time using a different fold for testing and the other folds for training. This method allows us to estimate the model's performance on unseen data.

For binary classification, we use accuracy and AUC (Area Under the Curve) performance metrics. Accuracy measures the proportion of correctly classified samples, while the AUC measures the ability of the model to distinguish between two classes, with 1 indicating perfect performance and 0.5 indicating random guessing.

For the regression models, we used two performance metrics: coefficient of determination (r^2) and predictive power score. The r^2 metric measures the proportion of variance in the target variable that can be explained by the model, with a value of 1 indicating a perfect fit and 0 indicating no correlation. The predictive power score (PPS) shows the ratio of how much better the model performed compared to a baseline (naïve) model, which always predicts the median of the target variable. The value of PPS ranges between 0 and 1 and it is defined as follows: $PPS = 1 - \frac{MAE_{model}}{MAE_{naïve}}$, where MAE is the Mean Absolute Error. For a great overview of evaluating ML models, we refer to the book of Zheng (2015).

3.3 Model intepretation

To gain insights into how our ML models make predictions, we utilized two techniques: built-in feature importance and SHAP (SHapley Additive exPlanations) values. The built-in importance metric is calculated based on how much the model's performance improves when that feature is included.

In addition to the built-in feature importance, we also used SHAP values, which is a state-of-the-art technique for model interpretation. SHAP values allow us to measure the contribution of each feature to an individual prediction. Here, we use SHAP values for the global interpretation of the model, namely, to see how the features affect the model prediction in general. To this end, we study how the SHAP values (impact on the prediction) change as the value of the feature varies from low to high. This plot is referred to as a SHAP summary plot that shows the contribution of the features for each student, where the feature names are on the y-axis and the x-axis shows the feature contribution/impact (SHAP value). For a comprehensive overview of the tools of interpretable ML, we refer to the book of Molnar (2020).

4 RESULTS

Predicting whether a student will choose a STEM major is a binary classification problem, where the value of the target variable is one if the major the student was most seriously considering when first entering postsecondary education after high school was in a STEM field, and zero otherwise. We predicted STEM major choice given that the student enters higher education. Thus, we excluded those students, who did not attend any college and the resulting data set contained 11,550 rows. We have tested multiple machine learning algorithms such as XGBoost, AdaBoost, and CatBoost, and on our data set the CatBoost algorithm achieved the highest performance. The mean cross-validated AUC score (i.e., the mean AUC on the five test sets resulting from the 5-fold-cross validation) is 0.801 (with a standard deviation of 0.007), moreover, the mean cross-validated accuracy of the model is 0.790 (with a standard deviation of 0.006). The results suggest, that it is possible to identify with relatively good accuracy which students will opt for a STEM major.

4.1 Features affecting STEM major choice

Besides evaluating the performance of the machine learning model, understanding its underlying mechanisms is critical for gaining insights into the factors driving its predictions. Namely, the goal of this section is to explore how the features influence the choice of a STEM major. Table 1 shows the top 10 most important features according to the built-in feature importance and SHAP values.

Table 1. The top 10 most important features in predicting STEM major choice. The features are ordered by the CatBoost importance, however, their rank according to the SHAP importance is written in parenthesis.

Variable	CatBoost's built-in importance	SHAP importance
Science ID (11th grade)	8.49	0.33 (2)
Sex	8.19	0.48 (1)
Science GPA	4.64	0.15 (5)
Math assessment (11th grade)	4.51	0.17 (4)
Math proficiency (11th grade)	4.28	0.15 (6)
Science for career	4.19	0.22 (3)
Math ID (11th grade)	3.41	0.13 (7)
Math theta score (9th grade)	3.04	0.08 (15)
English GPA	3.03	0.08 (11)
Math GPA	3.01	0.06 (21)

Source: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLS:09)

Table 1 suggests that the most important features are the students' science and mathematics identity, sex, mathematics skills, GPA scores (especially in science), and a binary variable that indicates whether they took a science course because they think they will need it for their career (*Science for career*). The science and mathematics identity variables are based on two other variables: one of them measures whether the students see themselves as a science/math person, while the other one measures whether they think that others see them as a science/math person. Naturally, we find

that the higher the value of the scale of science/math identity is the higher the model output is, i.e. the higher the probability of choosing a STEM major is. Hence, not so surprisingly, if high school students see themselves as science/math person, then they are more likely to opt for a STEM major in their university studies.

Furthermore, Table 1 suggests that sex also influences the students' decision to pursue a STEM major. Fig. 1 shows the SHAP summary plot of the top 20 most important features. From the figure, it is apparent that male students (when the value of *Sex* is low, i.e. 0) are more likely to choose a STEM major than females, which is in alignment with related works (Sahin et al. 2018; Vooren et al. 2022).

Besides the importance of science and mathematics identity, the figure also shows, that the higher the score in mathematics (assessment, proficiency, theta score) the higher the (positive) impact on the model's prediction (probability of choosing a STEM major). Interestingly, Figure 1 also suggests that the higher the GPA in English is the less likely that the student will decide to pursue a STEM major. One possible explanation for this phenomenon is that students who achieve high GPA scores in English may be more inclined to pursue liberal arts majors rather than STEM.

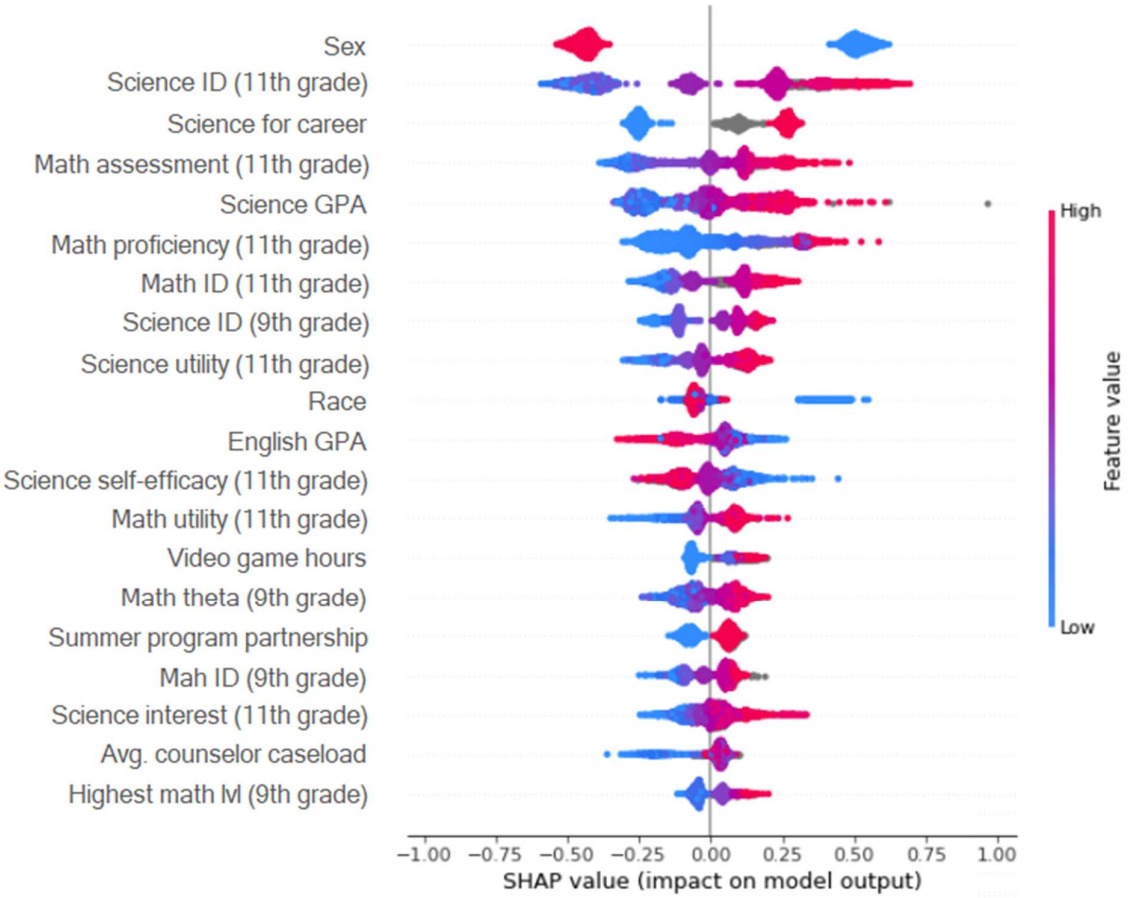

Fig. 1. SHAP summary plot of the 20 most important features affecting STEM major choice. One point is a feature's SHAP value for a student. Overlapping points are jittered to show the distribution of the SHAP values. The features are ordered by their importance. Source: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLS:09)

Finally, the reason why *Race* is also among the most influential variables in predicting students' decisions to pursue a STEM major is that Asian students are more likely (46%) to opt for a STEM major compared to other racial groups (20-25%), which is congruent with the findings of Sahin et al. (2018).

4.2 Predicting Science and Mathematics Identity

Our previous analysis predicted STEM major choices, and now we aim to understand the factors influencing students' science and mathematics identities, that are key predictors of STEM major choice. To this end, we trained two CatBoost regression models to predict the values of the scale of science and mathematics IDs in 11th grade, and thus, we excluded those variables that were assessed later on. On the other hand, here we do not filter those students that did not enter higher education, hence this analysis is based on a larger cohort, containing 19,940 rows for science identity prediction, and 20,020 rows[2].

To sum up, for predicting science identity we used the following attributes: *Sex*, *Race*, *Science for career* (takes science bec. needs it for career), *Science to be challenged* (takes science bec. likes to be challenged), *Science bec. does well* (takes science bec. does well in it), *Science can be learned* (agrees that most people can learn to be good at science), *Science self-efficacy (11th grade)*, *Science interest (11th grade)*, *Science utility (11th grade)*. Moreover, for predicting mathematics identity we considered the following variables: *Sex*, *Race*, *Math self-efficacy* (in 9th and 11th grades)*, Math interest (11th grade)*, *Math utility (11th grade)*, *More math bec. good at it* (plans to take more math courses because he/she is good at it), *Math to be challenged* (takes math bec. likes to be challenged), *Math bec. does well* (takes math bec. does well in it), *Math understanding frequency* (how often 9th grader thinks he/she really understands math assignments), *Algebra I* (final grade), *Math proficiency (11th grade)*, *Math assessment (11th grade)*, *Highest math lvl (9th grade)*. These variables were selected based on their correlation[3] with the math and science identity variables. The scale of students' science/mathematics interest, self-efficacy, and utility are composite variables created through principal component analysis, but we also study which subcomponents have the highest importance.

Our results show that the scale of students' science and math identities can be predicted relatively well. Specifically, the CatBoost regressor achieved r^2 values of 0.580 and 0.63 and yielded PPS of 0.392 and 0.423 for predicting science and mathematics identity, respectively. In what follows, we interpret the models to identify which students are most likely to develop science/math identities.

The effect of the variables in predicting science and mathematics identity is shown in Figure 2. The figure suggests that the most influential variables are the composite variables, i.e., self-efficacy, utility, and interest, and the *Science/Math bec. does well* non-composite variables. The most important subcomponents are the binary variables that indicate whether the student is enjoying math/science courses and/or taking

[2] We excluded those rows from the original data set where the science or math ID variable was missing.

[3] Pearson, Spearman correlation and predictive power score calculated with th ppscore Python package

math/science courses because they enjoy math/science – which are both incorporated into the science and math interest variables.

Naturally, the student's favorite subject is also a good predictor of science and math ID, since the favorite subject of these students is typically either science or mathematics. Besides the *Science/Math for career* variables, another important predictor of science/math identity, and hence of STEM major choice, is whether the student thinks that science or mathematics is useful for a future career – which are integrated into the science and math utility variables.

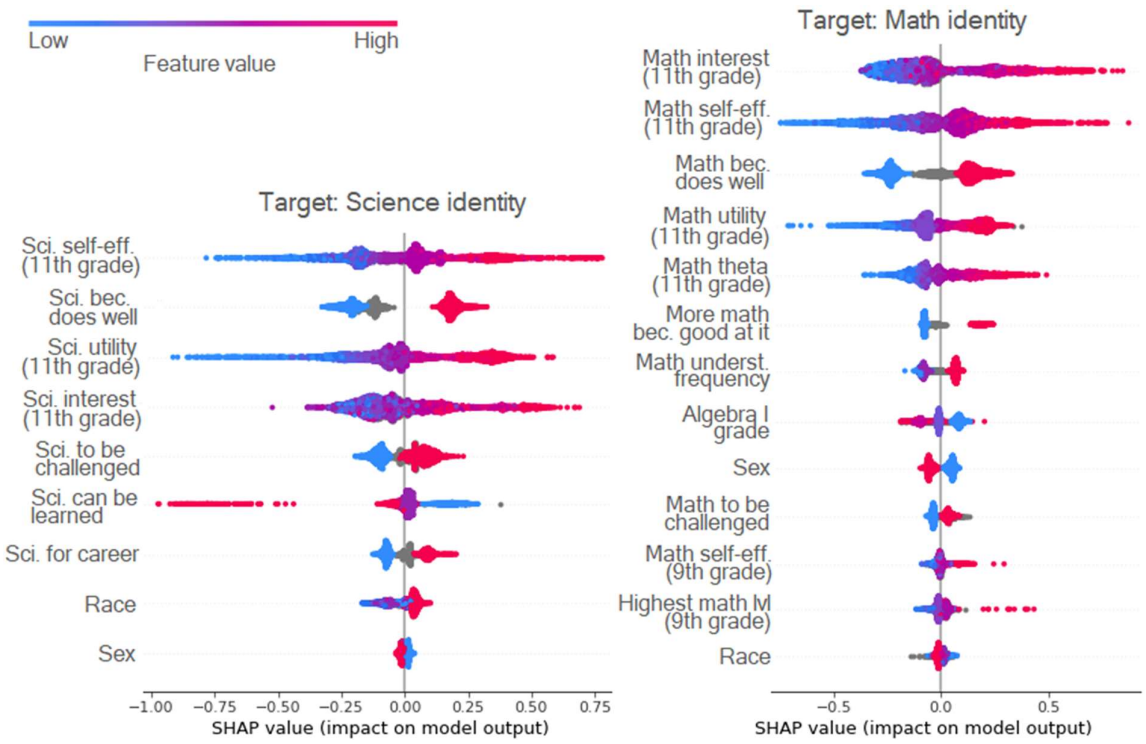

Fig. 2. SHAP summary plots for predicting science (left) and mathematics (right) identity. Source: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLS:09)

By comparing Figures 1 and 2, we can conclude that while gender appears to be a significant factor in students' decisions to pursue a STEM major, it is weakly associated with the students self-reported science or math identities. In other words, gender influences the decision to pursue a STEM major, however, it does not influence whether a student considers themself a science/math student.

5 SUMMARY AND CONCLUSION

This paper aims to investigate the predictability of students' choices in pursuing a STEM major and to identify the most influential factors in this decision-making process. Using machine learning models, we achieved relatively accurate predictions regarding which students are more likely to choose a STEM major. Sex, science or math identity, as well as scores and grades in math-related courses and tests, emerged as the most

crucial factors in predicting STEM major selection. Subsequently, our focus shifted towards understanding the determinants of science or math identity among students. Notably, while gender significantly impacted the decision to pursue a STEM major, it did not influence the identification as a science or math person. In other words, both boys and girls were equally inclined to be science or math individuals, yet girls were less likely to opt for a STEM major. The primary determinants of science or math identity included enjoyment of science or math courses, academic performance in these subjects, and the perceived usefulness of such courses for future career prospects. These findings contribute to a better understanding of the decision-making processes behind STEM major selection and science or math identity formation, offering valuable insights for policymakers and educators seeking to promote diversity and participation in STEM fields.

ACKNOWLEDGMENT

We thank Beata N. Johnson, Jonah Gerardus, and Tram Dang for their assistance with data analysis.

This material is based upon work by Main and Griffith supported by the National Science Foundation under Grant Number 2142697. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Marcell Nagy has been supported by the Fulbright Program and the doctoral student scholarship program of the Cooperative Doctoral Programme of the National Research, Development, and Innovation Fund. Roland Molontay is supported by The Rosztoczy Foundation and the National Research, Development, and Innovation Fund through the OTKA Grant PD-142585.

REFERENCES

Chang, Chi-Ning, Shuqiong Lin, Oi-Man Kwok, and Guan Kung Saw. 2023. "Predicting STEM Major Choice: a Machine Learning Classification and Regression Tree Approach." *Journal for STEM Education Research*: 1-17.

Grinsztajn, Leo, Edouard Oyallon, and Gael Varoquaux. "Why do tree-based models still outperform deep learning on typical tabular data?." In Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track. 2022.

Hastie, Trevor, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: Springer, 2009.

Kurban, Elizabeth R., and Alberto F. Cabrera. 2020. "Building readiness and intention towards STEM fields of study: using HSLS: 09 and SEM to examine this complex process among high school students." *The Journal of Higher Education* 91, no. 4: 620-650.

Molnar, Christoph. 2020. *Interpretable machine learning*. Lulu. com.

Sahin, Alpaslan, Adem Ekmekci, and Hersh C. Waxman. 2018. "Collective effects of individual, behavioral, and contextual factors on high school students' future STEM career plans." *International Journal of Science and Mathematics Education* 16: 69-89.

Vooren, Melvin, Carla Haelermans, Wim Groot, and Henriette Maassen van den Brink. 2022. "Comparing success of female students to their male counterparts in the STEM fields: an empirical analysis from enrollment until graduation using longitudinal register data." *International Journal of STEM Education* 9, no. 1: 1-17.

Wang, Xueli 2013. "Why students choose STEM majors: Motivation, high school learning, and postsecondary context of support." American Educational Research Journal 50, no. 5: 1081-1121.

Zheng, Alice. 2015. *Evaluating machine learning models: a beginner's guide to key concepts and pitfalls*. O'Reilly Media.