Predicting STEM Classroom Attrition Via Data Mining of Social Dynamics and Academic Metrics

James Squires, University of Georgia

James Squires is a 4th year honors mathematics and physics undergraduate attending UGA. He is interested in machine learning and AI. His primary research is in physics education research, with a focus on predicting at-risk students.

Additionally, James does research in neurosymbolic AI and theoretical physics. He is currently working on several other research projects involving AI theory, quantum mechanics, and computational physics.

Dr. Nicholas Young, University of Georgia

Nicholas Young is an assistant professor of physics with a focus on physics education research at the University of Georgia. He earned his PhD in physics and computational mathematics, science, and engineering at Michigan State University, and he was a postdoctoral scholar at the University of Michigan's Center for Academic Innovation. His current research focuses on assessment in introductory physics courses and graduate education in physics.

Dr. Nandana Weliweriya, University of Georgia

Nandana Weliweriya is a Senior Lecturer in the Department of Physics and Astronomy at the University of Georgia, Athens. With a background in physics education research, his teaching and research focus on physics problem-solving, active learning pedagogies, and undergraduate laboratory instruction.

Predicting STEM Classroom Attrition Via Data Mining of Social Dynamics and Academic Metrics

Work in Progress (WIP) Paper: Reducing attrition rates among STEM-enrolled college students is critically important to universities. Improved retention and higher graduation rates lead to better academic outcomes for both students and their respective institutions. Intervention methods to reduce attrition rates often prove less feasible in large classrooms without personalized student-teacher interaction. This work-in-progress paper examines whether social dynamics among students enrolled in introductory physics courses, combined with data encapsulating student grades, can predict at-risk engineering and non-engineering majors. Model performance is then compared to traditionally utilized variables. Key performance metrics are collected from the first third of the semester. These include grade data over the first exam, quizzes, and homework, along with responses to post-exam surveys. The surveys gather data on over a dozen variables including: demographic background, prior physics education, transfer status, motivation levels, grade expectations, and social dynamic variables like number of study peers, study time together, and study locations with peers. Both the survey and grade data of 238 students are then used to train logistic regression models designed to predict whether students will earn lower than a B–. Improved model fidelity for early semester predictions, when incorporating social variables, would indicate a promising direction for predicting educational outcome. In large classrooms, struggling students often go undetected by instructors; high-performing models would enable instructors to automatically identify at-risk students early enough that student intervention can positively impact academic outcomes like in future mid-term exams. This would allow instructors to direct support resources toward at-risk students more effectively. Improved support in engineering prerequisites like introductory physics and mathematics, where classroom sizes often exceed 50 students, can lower overall attrition and improve graduation rates. Thus, improving detectability of at-risk students in the classroom for engineering majors would increase the rate of engineering graduation and expand the future engineering workforce.

1 Introduction

The rapid growth of technological innovation characterized by the 21st century has created an increasing demand for STEM education. However, almost half of initially STEM-enrolled students either change to a non-STEM field or withdraw from the university, with this number drastically increasing for low-performing students [1]. Decreasing attrition rates among STEM-enrolled students would improve academic outcomes, thereby improving the available STEM workforce. Thus, it is pertinent that effective methods of decreasing attrition rates are developed.

Academic intervention is one such method of improving academic persistence [2]. However, large classroom sizes make detecting at-risk students highly challenging for teaching faculty. Therefore, recent initiatives have investigated the ability of machine learning models to identify at-risk students through the use of student data. Due to the highly complex profiles of students, currently achieved model performance likely remains suboptimal [3]. Thus, identifying new variables to further improve early prediction is a worthwhile research direction.

Due to the notorious reputation of introductory physics as a difficult course [4], we will specifically be investigating model employment in the context of Introductory Physics I.

Previous prediction models incorporate data that includes academic performance like course grades, academic background like entrance exams and prior GPA, demographic information like race and gender, behavioral traits like attendance, psychological traits such as motivation or self-efficacy, and background data such as family income [5, 6]. The most predictive variables are typically found to be exam grades, online learning patterns, and academic emotions [7]. Social dynamics largely remain under investigated in these studies.

This study focuses on several underutilized variables in the literature: the number of study peers a student has, the amount of time they spend studying with peers, their study locations, and their university transfer status. These variables are meant to capture social integration within an academic environment, specifically whether students fall through their academic social network, which has been shown to indicate low performance [8]. Logistic regression is used to perform a binary classification of at-risk and non at-risk students. A baseline model is built using grades, after which we analyze the change in model performance by introducing new variables; we assess both the predictive capabilities of social dynamic variables and more standard variables.

In this work-in-progress paper, we aim to answer the following research question:

- To what extent does the inclusion of student social study behaviors affect early-semester grade prediction models?

2 Background and Theory

2.1 Social Network Analysis and At-Risk Students

Social network analysis done in Stadtfeld et al. 2018 showed that socially isolated university students tend to obtain poorer academic outcomes than students that exist within social networks. Furthermore, students tend to form study partners with pre-existing friends [8]. Thus, determining whether a student studies with peers serves as a potential indicator of whether a student exists within a social network.

2.2 Repeated k-fold Cross Validation

If we want to improve the robustness of a model's performance metrics in a small dataset, we can partition the training dataset into k equally sized subsets. Then we use k - 1 of the partitions

for training data and the remaining set for testing performance. We then repeat this, using all k partitions as test data exactly once. We take the standard mean of performance metrics over k-trials. This is called k-fold Cross Validation.

k-fold Cross Validation allows models to more efficiently provide an estimate of performance on a small dataset, as all data is used for both testing and training.

To further reduce variance associated with the random partitioning of our data, we may then repeat the process by taking n k-fold partitions. We then take the standard mean of the n different k-fold performance metrics. This is known as repeated k-fold cross validation [9]. Additionally, we say k-fold regression is stratified when each partition has roughly equal proportions of positives to negatives.

2.3 Classification Tradeoff

There is a tradeoff between effectively classifying non at-risk (negatives) and at-risk students (positives). Most students are classified as not at-risk in a typical classroom dataset [10]. Because there are more examples of not at-risk students in the dataset, the model tends to perform better at identifying those students than at identifying at risk students.

This tradeoff poses a serious consideration for practical implementation in the classroom. If the amount of available resources to allocate are high, then we may desire that our model prioritize identifying at-risk students, which will inherently produce some false-positives (students who aren't at-risk who are said to be). Whereas, if available intervention resources are low, then we might pragmatically favor models that have high precision, leading to fewer false-positives, but also fewer interventions for truly at-risk students. It might also be unsafe to implement this model if the precision is not high enough, as we don't want to produce self-fulfilling prophecies at the detriment of otherwise non at-risk students [11].

3 Data and Processing

We collected data on 238 students over two semesters at a single institution. The data were collected from two Introductory Physics I courses. 22% of the data was collected from three sections of an active-learning, calculus-based introductory physics I course. The classroom sizes were about 72 students each. The other 78% was collected from two sections of a non-calculus-based introductory physics I course with lecture hall style classrooms and classroom sizes of about 210 students each. The data includes end of course grades as well as surveys given after the first exam.

All variables were converted into numerical values. We used one-hot encoding for variables with three or more discrete categories that could not be ordinally ranked (see Table 1). If the corresponding data-point of the student was not reported, then that student was dropped from the specific model data.

3.1 Model Variables

Here we describe the data we collected and any data processing steps that were taken before we included the data in our model. A table that describes how specific variables were numerically encoded into the models can be found in Appendix A.

Grades: We collected the following grade data from students: mid-terms, in-class grades, and homework grades. We also collected final course grades, which are classified as at-risk for students who have a C+ or below, and not at-risk for students who have a B– or greater. The cut-off was chosen at B– due to data availability; only 15% of our data has an at-risk classification. A lower cut-off would not provide enough data, whereas a higher cut-off might not represent truly 'at-risk' students. To reduce the dimensionality of our models, we take the averages of exam, homework, and in-class grades whenever applicable.

Gender: Gender was self-reported in our surveys. Self-reported gender was 99% binary, thus our dataset cannot reasonably support a non-binary column and we've chosen to omit non-binary gender data in the gender-model.

Race/Ethnicity: The dataset population was self reported at 65% white and 24% Asian. Other self-reported racial and ethnic identities were classified as Underrepresented Minority (URM) status. This made a binary variable in our model.

Academic Background: Students were asked what their highest level of prior physics was. This was encoded as an ordinal variable. We also assess student transfer status: whether students transferred into their current institution was encoded as a binary variable. Physics background is meant to probe student preparation for course material, while transfer status might indicate alternative academic preparation.

Social Dynamics: We collected three social variables through a survey distributed within a week after the first exam. Students self-reported: the number of peers they study with, the amount of hours they study with peers, and the locations they study with peers at. We encoded the number of peers into a binary category: either 0 or 1 or more study peers. The self-reported study locations with peers were heuristically categorized. Four common categories were produced: academic buildings such as libraries, homes or dorms, dining halls or cafes, and other.

Motivation and Expectations: We asked students their motivation to succeed in this course relative to other courses based on a Likert-scale. We also asked the students what grade they expect to receive. The student's predicted grade is then turned into a binary variable based on the same at-risk criterion.

3.2 Model Creation and Evaluation

To evaluate the performance of the logistic regression models, we repeated stratified 5-fold cross validation 10 times. We used stratification to ensure representative subsamples. Students missing one or more variables were dropped from each dataset.

Our baseline model uses three variables: exam 1, homework grades, and in-class grades. We iterate over the baseline logistic regression models by adding each variable to a baseline model and running the baseline model against the extended 4 variable model. A lift or change (Δ) of 0 is the default measurement of the baseline model metric performance. A positive lift represents an improvement in extended model performance. We assess 11 different variables against the baseline model.

4 Results

The baseline model demonstrated a recall of 0.594 ± 0.168, a precision of 0.337 ± 0.096, an accuracy of 0.768 ± 0.052, and a Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of 0.808 (see Appendix C). 85% of the students were classified as non at-risk. For comparison, a similar logistic regression model obtained an ROC-AUC of 0.88 and an accuracy of 0.78 on 915 students [10].

The inclusion of self-reported studying in an academic building (with peers) significantly improved model performance with an ROC-AUC improvement of 0.023. The recall and F1-scores also improved, suggesting the variable improved prediction of at-risk students (see Appendix C). The highly negative, standardized coefficient of $\beta = -0.612$ corresponding to the academic building parameter suggests that students who study in academic buildings such as libraries (with peers) are less likely to be considered at-risk compared to students who self-reported not studying in academic buildings with peers. Note that this category treats students who study with no peers the same as students who study in non-academic buildings.

While other social variables such as having study peers or high social study hours demonstrate improved performance metrics, the ROC-AUC lift is not statistically significant (see Fig 1). However, the extremely negative coefficients found in Appendix C imply that social behavior around course material is associated with lower odds of being classified as at-risk.

5 Limitations

This dataset does not include students who withdrew from the courses. This likely leads to a survivorship bias in which many of the students who would otherwise be classified as at-risk early into the semester were not included in our assessment. Additionally, this dataset was only collected from two instructors at a single institution across two semesters, limiting the generalizability of the results.

Logistic regression only captures exponential relationships between odds and a given variable; thus, other nonlinear relationships might go undetected. Alternative models such as random forest regression might capture these relationships.

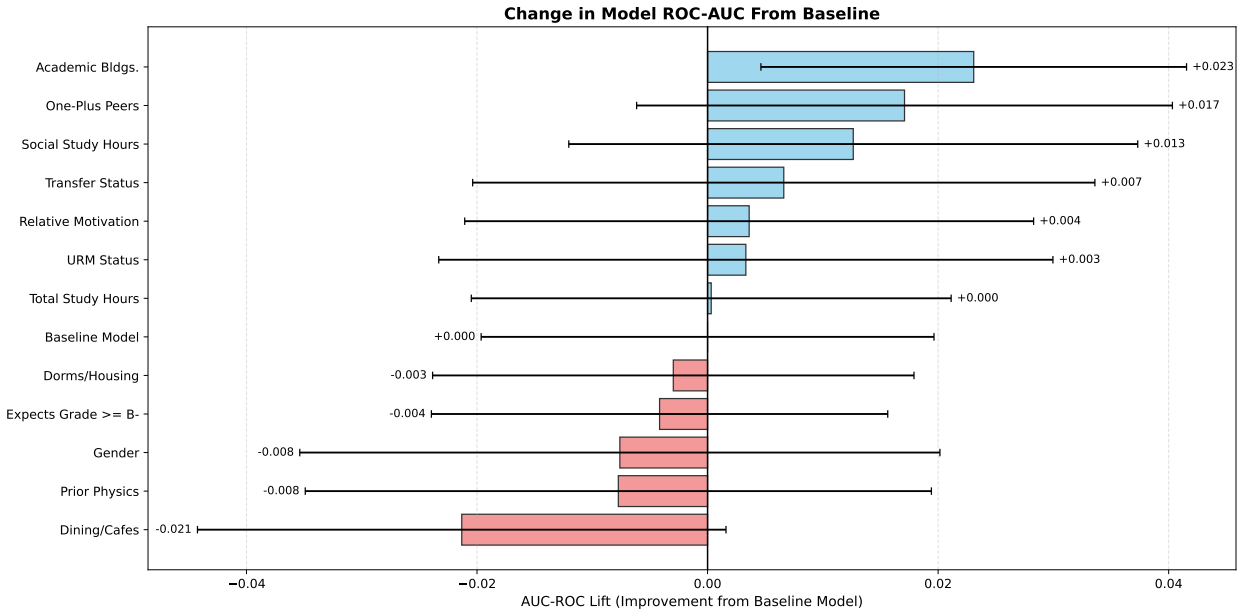

Figure 1: ROC-AUC lift from baseline models across 10 repeats of 5-fold cross-validation. Error bars represent standard error for 50 fold evaluations $\pm 1\sigma/\sqrt{10}$.

6 Conclusion and Future Expectations

Prior literature suggests students who fall outside of peer social-networks are more likely to perform poorly in their classes. We aimed to capture this relationship in grade prediction models and investigate whether the new variables improve model performance. The results indicate that social dynamics indicating belonging in a social-network, indeed, have predictive capabilities. Inclusion of whether students study alone, how long they study with peers, and whether they study in academic buildings like libraries, all improved model performance and had parameter coefficients with negative associations between social-network belonging and at-risk classification. This suggests a promising avenue for predicting at-risk students through the use of social-network analysis. By capturing social-dynamics amongst peers, models incorporate another dimension of student behavior that is not easily captured through grades or academic background.

Next steps will include model refinement. We will also continue to expand the dataset and include more classrooms in order to improve the robustness of our analysis. Once this is done, we intend to apply these prediction models in large classrooms to determine the impact machine learning informed intervention has in practice.

7 References

[1] X. Chen, "Stem attrition among high-performing college students: Scope and potential causes," *Journal of Technology and Science Education*, vol. 5, 03 2015.

[2] J. M. Harackiewicz and S. J. Priniski, "Improving student outcomes in higher education: The science of targeted intervention," *Annual Review of Psychology*, vol. 69, pp. 409–435, 01 2018.

[3] Y. A. Alsariera, Y. Baashar, G. Alkawsi, A. Mustafa, A. A. Alkahtani, and N. Ali, "Assessment and evaluation of different machine learning algorithms for predicting student performance," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–11, 05 2022.

[4] F. Ornek, W. R. Robinson, and M. P. Haugan, "What makes physics difficult?," *International Journal of Environmental and Science Education*, vol. 3, p. 30–34, 01 2008.

[5] I. Issah, O. Appiah, P. Appiahene, and F. Inusah, "A systematic review of the literature on machine learning application of determining the attributes influencing academic performance," *Decision Analytics Journal*, vol. 7, p. 100204, 06 2023.

[6] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student' performance prediction using machine learning techniques," *Education Sciences*, vol. 11, p. 552, 09 2021.

[7] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Applied Sciences*, vol. 11, p. 237, 12 2020.

[8] C. Stadtfeld, A. Vörös, T. Elmer, Z. Boda, and I. J. Raabe, "Integration in emerging social networks explains academic failure and success," *Proceedings of the National Academy of Sciences*, vol. 116, pp. 792–797, 12 2018.

[9] S. S. Bates, T. Hastie, and R. Tibshirani, "Cross-validation: what does it estimate and how well does it do it?," *Journal of the American Statistical Association*, vol. 13, pp. 1–22, 04 2021.

[10] C. Zabriskie, J. Yang, S. DeVore, and J. Stewart, "Using machine learning to predict physics course outcomes," *Physical Review Physics Education Research*, vol. 15, 08 2019.

[11] L. Jussim and K. D. Harber, "Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies," *Personality and Social Psychology Review*, vol. 9, pp. 131–155, 05 2005.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical learning, Second Edition : Data mining, inference, and Prediction*. Springer, 2 ed., 2009.

Appendix A: Table of Variables with Numerical Representations

Table 1: Variables used in baseline and extended predictive models

Variable	Description	Numerical Representation
Final Grade (Target)	Whether the student achieved a B– or above	Binary: 1 (<B–), 0 otherwise
Exams	Mid-term assessments (10–25% of total grade)	Continuous: [0, 1]
Homework Average	Average of homework assignments	Continuous: [0, 1]
In-Class Average	Average in-class participation grade	Continuous: [0, 1]
Physics Class	Highest physics course completed (None, HS, AP/College)	Ordinal: 0, 0.5, 1.0
Study Hours	Self-estimated average weekly study hours	Continuous: [0, 1]
Relative Motivation	Motivation relative to other courses (Likert scale)	Ordinal: (0, 0.25, 0.5, 0.75, 1)
Expected Grade	The student expects a B– or above	Binary: 1 (Yes), 0 (No)
Transferred	Has the student transferred to this university	Binary: 1 (Yes), 0 (No)
Social Study Hours	Weekly study hours with peers	Continuous: [0, 1]
One-Plus Peers	Self-estimated number of study peers	Binary: 1 (Yes), 0 (No)
Study Location	Self-reported study locations with peers	One-hot encoded
Gender	Self-identified gender	Binary: 1 (Male), 0 (Female)
URM Status	Self-identified race/ethnicity	Binary: 1 (URM), 0 (white or Asian)

Appendix B: Metrics Formulas

Table 2: Definitions of Performance Metrics for At-Risk Classification

Metric	Formula	Description
Accuracy	$(TP + TN)/(TP + TN + FP + FN)$	Measures the proportion of correct classifications
Recall	$TP/(TP + FN)$	The ratio of at-risk students correctly identified as at-risk
Precision	$TP/(TP + FP)$	Measures the amount of noise in at-risk student predictions
F1-Score	$2TP/(2TP + FP + FN)$	The harmonic mean of precision and recall

Appendix C: Full Model Results

Table 3: Standard mean ± standard deviation (SD) of performance metrics across stratified 10 repeated 5-folds

Variable	Recall ± SD	ROC-AUC ± SD	Precision ± SD	Accuracy ± SD	F1 ± SD
Academic Buildings	0.651 ± 0.169	0.829 ± 0.058	0.393 ± 0.090	0.800 ± 0.046	0.481 ± 0.093
Relative Motivation	0.625 ± 0.183	0.803 ± 0.078	0.371 ± 0.113	0.782 ± 0.062	0.455 ± 0.121
Dorms/Housing	0.592 ± 0.204	0.803 ± 0.066	0.339 ± 0.097	0.770 ± 0.058	0.421 ± 0.115
Transfer Student	0.565 ± 0.228	0.801 ± 0.085	0.356 ± 0.131	0.792 ± 0.057	0.431 ± 0.158
Social Study Hours	0.627 ± 0.202	0.831 ± 0.078	0.359 ± 0.119	0.775 ± 0.067	0.449 ± 0.136
Dining Halls/Cafes	0.563 ± 0.193	0.785 ± 0.072	0.319 ± 0.078	0.763 ± 0.049	0.399 ± 0.099
Prior Physics	0.555 ± 0.214	0.786 ± 0.086	0.312 ± 0.107	0.761 ± 0.056	0.394 ± 0.134
One-Plus Peers	0.621 ± 0.179	0.827 ± 0.073	0.373 ± 0.108	0.799 ± 0.051	0.459 ± 0.120
URM Minority	0.553 ± 0.190	0.801 ± 0.084	0.324 ± 0.096	0.778 ± 0.043	0.404 ± 0.117
Gender	0.528 ± 0.215	0.785 ± 0.088	0.298 ± 0.107	0.754 ± 0.057	0.373 ± 0.132
Total Study Hours	0.568 ± 0.189	0.812 ± 0.066	0.333 ± 0.121	0.763 ± 0.069	0.410 ± 0.129
Expects Grade ≥ B-	0.560 ± 0.185	0.804 ± 0.063	0.325 ± 0.104	0.764 ± 0.054	0.402 ± 0.116
BASELINE (Grades)	0.594 ± 0.168	0.808 ± 0.062	0.337 ± 0.096	0.768 ± 0.052	0.423 ± 0.106

Table 4: Model coefficients and change in performance metrics (Δ) relative to baseline.

Variable	Coefficients	Recall Δ	ROC-AUC Δ	Precision Δ	Accuracy Δ	F1 Δ
Academic Buildings	-0.612	+0.101	+0.023	+0.078	+0.037	+0.089
Relative Motivation	0.406	+0.048	+0.004	+0.039	+0.022	+0.040
Dorms/Housing	0.267	+0.042	-0.003	+0.024	+0.008	+0.030
Transfer Student	0.358	+0.016	+0.007	+0.051	+0.035	+0.044
Social Study Hours	-0.867	+0.013	+0.013	+0.004	-0.000	+0.006
Dining Halls/Cafes	-0.148	+0.013	-0.021	+0.004	-0.000	+0.007
Prior Physics	-0.052	+0.006	-0.008	+0.007	+0.004	+0.008
One-Plus Peers	-0.437	+0.003	+0.017	+0.001	+0.005	+0.003
URM Status	0.282	-0.007	+0.003	+0.012	+0.010	+0.007
Gender	-0.005	-0.014	-0.008	-0.003	+0.001	-0.006
Total Study Hours	0.172	-0.026	0.000	-0.015	-0.007	-0.020
Expects Grade ≥ B-	0.051	-0.034	-0.004	-0.013	-0.004	-0.021
BASELINE (Grades)	—	0.000	0.000	0.000	0.000	0.000

To maintain model integrity, students missing data for a particular variable were excluded from that specific model's dataset. This new dataset is calculated against its own baseline. The Baseline variable found at the bottom of tables C and D contains all 238 students.

Appendix D: Logistic Regression

Binary logistic regression is a machine learning technique used to predict the classification of binary data or *targets* $Y_i \in \{0, 1\}$ for all data samples i given corresponding data points $\mathbf{x}_i \in \mathbb{R}^n$, where n variables or *parameters* are considered. The curve produced by optimizing maximum likelihood estimation is a sigmoidal function:

$$\sigma(\mathbf{x}) = \frac{1}{1 + \exp\{-(\beta^T \mathbf{x} + \beta_0)\}} \tag{1}$$

where our coefficients or *weights*, stored in a weight vector β, along with our intercept β_0, are produced from the optimization on our given training data. $\sigma(\mathbf{x})$ represents the probability $P(1|\mathbf{x}_i) \in [0, 1]$ of a classification given a data-point $\mathbf{x}_i$. We choose a threshold $\tau \in [0, 1]$ for our estimate classification $\hat{Y}$ on future data:

$$\hat{Y} = \begin{cases} 1 & \text{if } \sigma(\mathbf{x}) \geq \tau \\ 0 & \text{if } \sigma(\mathbf{x}) < \tau \end{cases} \tag{2}$$

Logistic regression is useful for its interpretability. The coefficients $\beta_1, ..., \beta_n$ indicate the relationship between a parameter and the outcome. A large positive coefficient $\beta_k \in \mathbb{R}$, $1 \leq k \leq n$, indicates that, as a parameter $x_k \in \mathbb{R}$ increases, the odds of $Y_i = 1$ also increase [12].