Real-World Learning Analytics: Modeling Student Academic Practices and Performance

Chantal D. Larose[1], Kim Y. Ward[1]
[1]Eastern Connecticut State University, 83 Windham Street, Willimantic CT 06226

Abstract
We present a real-world learning analytics investigation to model student academic practices and performance in foundational mathematics courses. Segmentation analyses seek to clarify patterns through modeling subpopulations of student academic practices and various common course requirements. The effects of cost matrices and of rebalancing the data are examined, and their impact on the conclusions quantified. Final results include data driven guidelines for future student interventions.

Key Words: learning analytics, data science, cost matrices, predictive analytics, modeling

1. Introduction

Eastern Connecticut State University (ECSU) is a small liberal arts university, comprised of about 5,000 students. ECSU is continually working on increasing the number of students who successfully complete their first year of college. For the Mathematical Sciences Department, this means helping students pass courses in the Math Foundations Program. The ECSU Math Foundations Program includes:

- MAT 099 - Algebra Essentials
- MAT 135P - Math for Liberal Arts Plus
- MAT 155P - PreCalculus Mathematics Plus

Courses have a common syllabi, tutoring requirements, and final exams. Tutoring is done in the Mathematics Achievement Center (MAC), the University's math tutoring center.

This project marks the first quantitative look at how we can help students succeed. We use learning analytics to uncover the relationship between students' academic practices and their subsequent performance. Here, “Practices" refers to students attending tutoring, while “Performance" refers to students’ midterm and final course grades.

There are course requirements to be aware of. Students need 540 minutes (9 hours, or about 1.3 hours per week) in the MAC tutoring center before midterm grades are assigned. Students need a total of 1,080 minutes (18 hours) in the MAC before finals course grades are assigned.

We want to quantify the correlation and patterns between (i) Midterm exam grades and midterm tutoring attendance, and (ii) Final course grades with midterm exam grades, midterm tutoring attendance, and final tutoring attendance.

2. Exploratory Data Analysis

Before we jump into modeling, we need to examine the behavior between our target variables of interest and their predictors: (1) The target Midterm Grade and the predictor Midterm MAC Minutes, and (2) the target Final Course Grade and the predictor Final MAC Minutes.

2.1 Target Variable: Midterm Grade

Figure 1 shows a normalized histogram of Midterm MAC Minutes, with each bar colored by the distribution of Midterm Grades of the students in that bar. The vertical line indicates the amount of MAC tutoring time a student is required to attend by midterms. There is a pronounced increase in the proportion of A's and B's in each bar as we progress across the chart from Midterm MAC Minutes of zero toward the required time of 540 minutes. Thus, we anticipate Midterm MAC Minutes being important to the prediction of Midterm Grade.

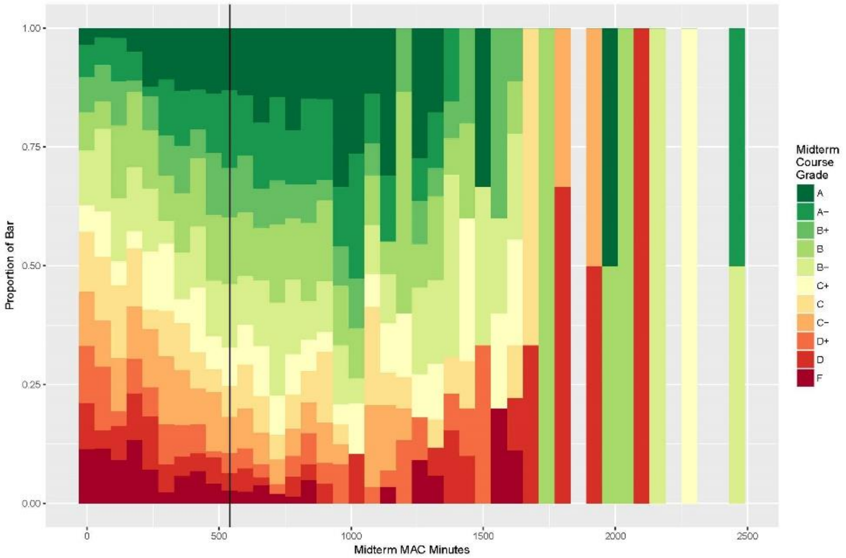

Figure 1: Normalized histogram of Midterm MAC Minutes with an overlay of Midterm Grade. The vertical line represents the required tutoring time.

2.2 Target Variable: Final Course Grade

Now we examine the relationship between the target variable Final Course Grade and the predictor variable Final MAC Minutes. Figure 2 shows a normalized histogram of Final MAC Minutes with an overlay of Final Course Grade. We see a similar pattern to that of Figure 1, namely that students perform better the more they approach the tutoring requirements. We anticipate Final MAC Minutes being important in predicting Final Course Grade.

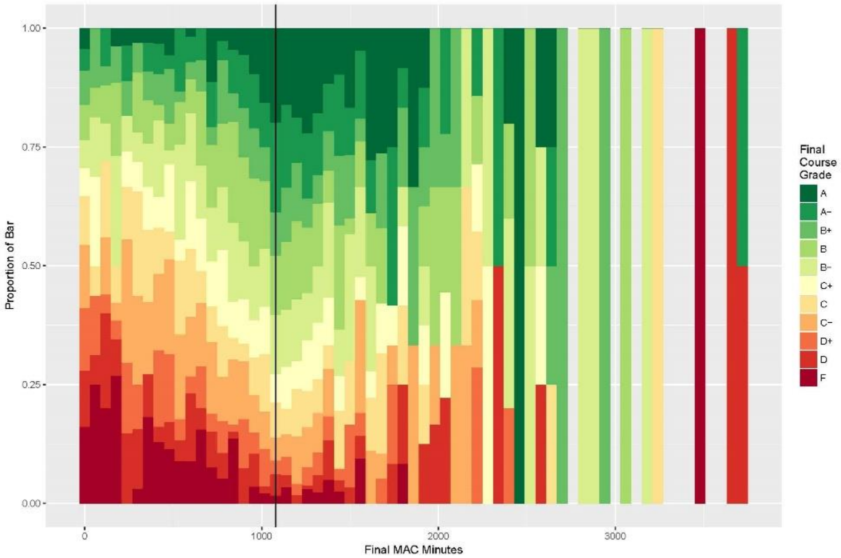

Figure 2: Normalized histogram of Final MAC Minutes with an overlay of Final Course Grade. The vertical line represents the required tutoring time.

To alleviate the issue of double-counting Midterm MAC time when considering Final MAC time, we create a new variable: *Extra MAC Minutes*, which looks at the additional time students went to tutoring after midterms.

3. CART Modeling

3.1 Classifying Midterm Grades

We apply classification and regression tree (CART) modeling (Breiman et al, 1984; Therneau et al, 2017) to the target variable Midterm Grade and predictor variable Midterm MAC Minutes. The model finds a single split to optimize the purity and balance of the terminal nodes: Midterm MAC minutes below 304 minutes, versus Midterm MAC minutes at or above 304 minutes.

Of the students who achieved a Midterm MAC time of at least 304 minutes (about 77% of the data set), 20% earned an A for a midterm grade, and only 4% earned an F for a midterm grade. Of the students who did not achieve at least 304 minutes of tutoring time (about 23% of the data), only 4% of these students earned an A for a midterm grade, and 20% earned an F. The model suggests that students who attend the MAC at least 304 minutes, or about 5.1 hours, before midterms have a much greater chance of receiving an A grade compared to students who do not reach that time commitment.

3.2 Classifying Final Course Grades

There are three subgroups within our student population: (1) Students who do not attend tutoring at all by midterms; (2) Students who attend tutoring before midterms but fall short of the required time; (3) Students who meet or exceed the midterm tutoring requirement. The three groups have very different shapes, centers, and variation. Therefore, moving forward, each group will be modeled separately. This will allow us to better capture and describe the different groups' behavior. Additionally, we collapse the separate letter grades into a binary "Pass or Fail" target variable.

3.2.1 Students with No Midterm MAC Time

The CART model for predicting the final course grade of students with no midterm tutoring found that the two most important variables when classifying students with no tutoring attendance is Midterm Grade and Extra MAC Time. Of the students who received a C or better on the midterm (~40% of the data), about 97% passed the course. For the remaining students, those who went to 702 or more minutes of tutoring after midterms passed about 68% of the time, while students who did not go for at least 702 minutes passed the course only 35% of the time.

A contingency table of predictions versus the real values is given in Table 1. Using these values, we see that the model for students with no MAC tutoring has an accuracy of 86.8%, sensitivity of 97.6%, and specificity of 63.2%.

Table 1: Contingency table for CART Model of Students with No MAC Time

	Prediction	
Reality	*Pass*	*Fail*
Pass	41	1
Fail	7	12

3.2.2 Students with Minimal MAC Time

The CART model for students with positive but less than the required tutoring time found that the two most important variables for predicting Final Grades are, again, Midterm Grade and Extra MAC Time. Students with a C or better on the Midterm pass 95% of the time. Students with an F on the Midterm fail 83% of the time. For students between these groups, Extra MAC Time of 772 minutes is the threshold between passing 95% of the time and passing only 50% of the time. After these two variables, pre-existing Midterm MAC Minutes further refine the classification.

A contingency table of predictions versus the real values is given in Table 2. Using these values, we see that the model for students with minimal MAC tutoring has an accuracy of 85.9%, sensitivity of 92.3%, and specificity of 58%.

Table 2: Contingency table for CART Model of Students with No MAC Time

	Prediction	
Reality	*Pass*	*Fail*
Pass	276	23
Fail	29	40

3.2.3 Students with Sufficient MAC Time

The CART model for students who met or exceeded the required MAC time (not shown) had a specificity of 29.6%. Therefore, we:

1. Rebalance the data, so 30% of the data set who met the MAC requirement failed the course (up from 10%).

2. Weight the errors differently; a false positive is five times as bad as a false negative.

After these changes are applied, we obtain the contingency table shown in Table 3. The model now has an accuracy of 78.5%, sensitivity of 82.9%, and specificity of 68.9%. Since sensitivity is the most important metric for the goals of this project, this model is satisfactory.

Table 3: Contingency table for CART Model of Students with Sufficient MAC Time

	Prediction	
Reality	*Pass*	*Fail*
Pass	345	71
Fail	60	133

4. Conclusions and Future Work

This project represents the first step toward quantifying how to help students through the Math Foundations Program at ECSU. Depending on Midterm tutoring attendance and exam performance, we can identify students who need intervention to perform their best in the course. Rebalancing and unequal misclassification costs help us zoom in on students who need help the most.

Conclusions are currently limited by not having student GPAs before and after the semester in which they take the course. Next steps include addressing missing data (< 8% of records) and more model-tweaking. Other areas of interest include separating data out by class year and course number, to further refine the model predictions.

Acknowledgements

This project is supported in part by the CSU-AAUP 2018-2019 University Research Grant. Travel to present this project at JSM 2018 was supported in part by the Eastern Connecticut State University's 2017-2018 Faculty Development Grant.

References

Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. Classification and Regression Trees. Chapman & Hall/CRC Press, Boca Raton FL, 1984.

Daniel T. Larose and Chantal D. Larose, Data Mining and Predictive Analytics, Second Edition. Wiley, 2015.

Terry Therneau, Beth Atkinson and Brian Ripley (2017). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-11. https://CRAN.R- project.org/package= rpart