Using Statistical Analysis and Machine Learning to Analyze Student Performance in Math 1020/1021

Cayden Huffman, Kennedy Jeffery, Devin Johnson, Jacob Johnston, Faris Khattak, Haoxiang Lin, Bayla Patin

Department of Mathematics, Louisiana State University,
Advisors: Nadejda Drenska and Christian Ennis

Introduction

- Math 1020/1021 is College Algebra with Corequisite material. In Fall 2022, these were treated as a single 5-hour credit course. By 2024, these two courses were treated as a 2- and 3-hour credit course, respectively, which are taken simultaneously.
- We seek to determine whether one model was more effective than another, as well as identify common issues in student performance.
- Finally, we sought to speculate student grades using classification and regression models, without the use of qualitative data. [1]
- These goals were motivated by an effort to identify at-risk students, who are most in need of academic intervention, early in the semester [2].
- This study was conducted without the use of any qualitative data, using only grades from these semesters.
- This study was approved by the LSU Institutional Review Board (IRB).

Data and Methodology

The grades of students from Fall 2022 and 2024 were given by tabular data sets composed of 806 and 705 rows, respectively. Using pandas and Excel, we generated plots and counts reflecting relations between grade item categories and student performance. Using the Math 1021 grading scale, we generated Figure 1, a time series of student performance given different categories.

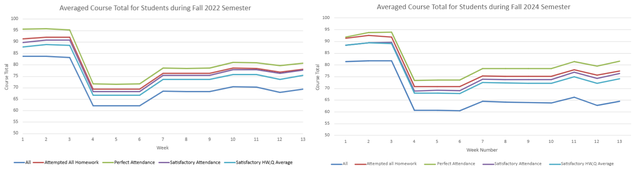

Figure 1. Time series charts of averaged student grades under different criteria for the Fall 2022 and 2024 semesters.

- Important Weeks: Test 1 entered on Week 4, Test 2 entered on Week 7, Test 4 entered on Week 12. Test 3 was entered on Week 10 for 2022 and on Week 11 for 2024.

Statistical Analysis

- We examine the distribution of averages given if a student passed or failed Math 1021 to gain insight on the trends noticeable in both sets of students.

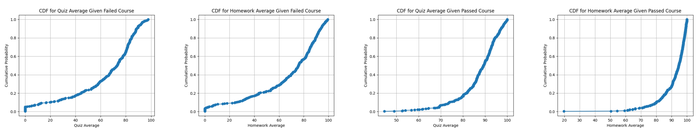

Figure 2. CDFs of conditional distributions given pass or failure for Fall 2024 data.

Insights from Counts

- Using pandas and Excel, we gathered counts from the data, and obtain useful insights on student performance upon applying Bayes' Theorem. This gives Table 1.
- HW, Q, CP, and LP mean homework, quiz, class participation, and lab participation averages respectively. These averages are computed at the end of the semester.

Given Event	Probability of Failing	
	Fall 2022	Fall 2024
HW < 70	0.875	0.894
Q < 70	0.801	0.846
HW,Q < 70	0.916	0.926
CP < 70	0.729	0.677
LP < 70	0.612	0.783
CP,LP < 70	0.796	0.831
HW,Q,CP,LP < 70	0.984	0.972

Table 1. Probability a student will fail given certain events. Separated by semester.

Visualizations

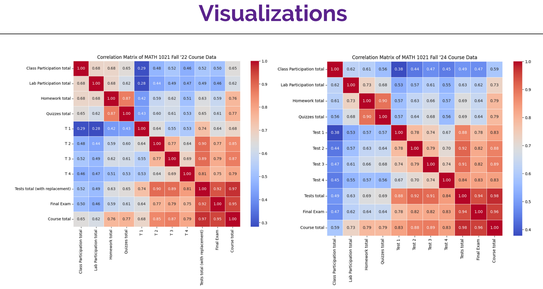

Figure 3. Heatmaps between grade categories for the Fall 2022 and 2024 semesters.

- We see general correlations remain similar between semesters with some notable changes.
- The non-test items with the highest correlation are homework and quiz averages.

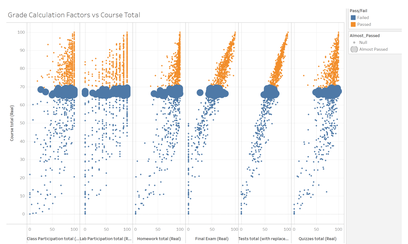

Figure 4. Fall 2022 scatter plots of grade factors against course total.

Using Machine Learning for Classification

We sought to develop an ML model to classify students as pass or fail based on a small amount of data (items up to and including the first test). We tested on future data as well to compare accuracy. Ideally, the model should minimize false pass predictions over maximizing true failure predictions. Highest accuracy was obtained with logistic regression.

Metrics	At Test 1	At Test 2	Up to Final
Accuracy	87%	88%	90%
Failure Accuracy	81%	80%	88%
Pass Accuracy	92%	96%	92%

Table 2. Metrics of the logistic regression model applied to combined Fall 2022 and 2024 data.

- Only marginal improvement when adding in a larger amount of data available in the semester.

Using Regression for Score Speculation

Multiple regression models were tested, with maximum accuracy being obtained with multiple linear regression. Using data up to the first test and up to the second test, we obtain higher accuracy and lower error metrics. To increase accuracy, we combined the two data sets. We saw this as valid since similar behaviors are seen between the two semesters.

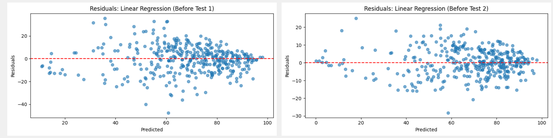

Figure 5. Residual plots for MLR using data up to Test 1 and Test 2.

Metrics	Values	
	Up to Test 1	Up to Test 2
Pass/Fail Accuracy	$78.34\% \pm 1.91\%$	$85.55\% \pm 1.59\%$
MAE	9.61 ± 0.4	6.35 ± 0.25
MSE	159.79 ± 13.68	73.11 ± 6.38
RMSE	12.63 ± 0.54	8.54 ± 0.37
R^2	0.6711 ± 0.0334	0.8494 ± 0.0161

Table 3. Averaged metrics across 100 seeds separated by amount of data used.

Acknowledgements

- We would like to thank Prof. Nadejda Drenska for acting as PI for our study and allowing this study to be conducted.
- We thank the Math 1021 administrative team, especially Stephanie Kurtz and Phoebe Rouse as this project would have been impossible without their approval and support.
- We thank the Math 1021 students of the past, present, and future as our motivation is due to you all.

References

1. Korchi, A., Messaoudi, F., Abatal, A. et al. Machine Learning and Deep Learning-Based Students' Grade Prediction. *Oper. Res. Forum* 4, 87 (2023).
2. Boretz, E. Midsemester academic interventions in a student-centered research university. *Journal of College Reading and Learning*, 42(2), 90–108 (2012).