UNIVERSITY OF SOUTH CAROLINA BEAUFORT

Machine Learning for Students Math Course Performance

Christoph Hagenauer,

Faculty Mentor Dr. Davide Fusi

Computational Science / Mathematics, University of South Carolina Beaufort, Bluffton, SC 29909

Abstract

This project is to analyze the success of students in their first ever Math course at USCB. For this purpose, we have a few features for each student. We are able to use the high school math courses they took, including the specific subject and class type (regular, honors, AP, …). To also use the high school they attended, we used US News' ranking of high schools. This includes the overall experience of a student (Facilities, student to faculty ratio, Cafeteria, …) and the university readiness score (accounts for AP classes, and how graduates perform in university). Additionally, we factor in how they performed compared to their respective class and their ACT / SAT score. In combination with the students first ever Math course at USCB we perform supervised Machine Learning in order to build a model. We hope to be able to use this model in the future to determine which class incoming Freshman / Transfer Students should take first at USCB.

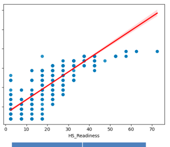
Methods

- Started with dataset provided by university examining the data
- Added High School Score and High School Readiness score from US News (only public high schools were considered, and also make up the majority of students at USCB)
- Converted ACT score to SAT score, however later build separate models for both
- Reduced used class to only in-person classes because grade distribution of online classes implicated biased data
- Merged student who took their first class in the Fall and in the Spring into common column
- Replaced grades with Pass or Fail (at USCB a grade of C or higher is required in order to be considered passing)
- Data visualization to get a better understanding of the data
- Scaled data
- Applied Multiprocessing algorithm to search best features for the model
- Build various models
- Created interactive Dashboard

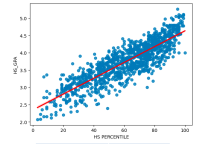
Glance at Data

SEMESTER	YEAR	COURSE	COURSE GRADE	HS PERCENTILE	HS GPA	SAT SCORE	ACT SCORE	HS Readiness	HS Score	PASS / FAIL
Fall	2017	MATHB111	F	54	2.925		15	7.5	27.5	0
Fall	2017	MATHB111	B	44	3.208	430		32.5	87.5	1
Fall	2017	MATHB111L	A	87	3.89	530		17.5	57.5	1
Fall	2017	MATHB101	W	74	3.669	390		17.5	27.5	0
Fall	2017	MATHB111	D+	0	2.28	500		27.5	72.5	0
Spring	2017	MATHB101	B+	67	3.093	450		2.5	27.5	1
Fall	2017	MATHB111L	A	24	3.83	540				1
Fall	2017	MATHB111	A	92	4.658		24	2.5	42.5	1
Spring	2017	MATHB111	A	86	4.16	540	24	17.5	77.5	1
Fall	2017	MATHB221	F	27	3.22		17	2.5	37.5	0

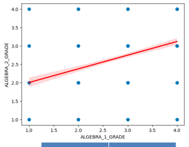
Selected Plots to Identify columns



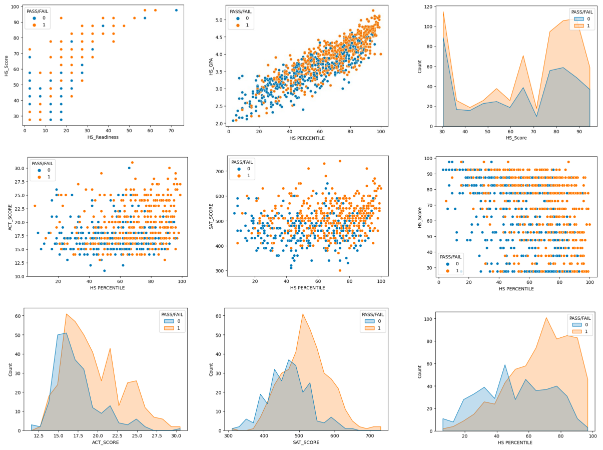
Correlation	P-Value
0.8103	$1.1 * 10^{-289}$

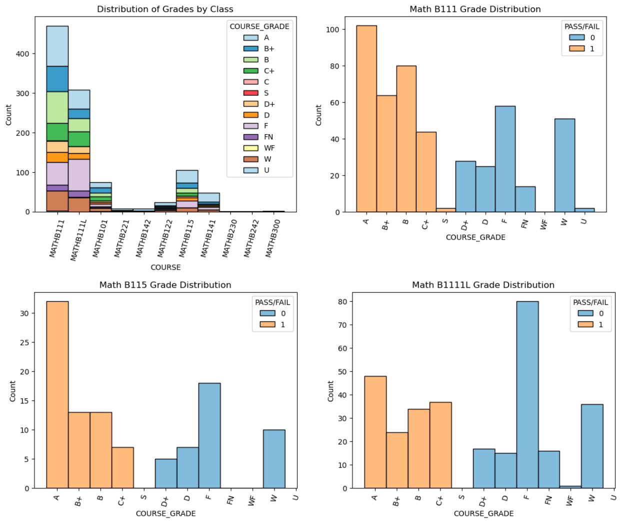


Correlation	P-Value
0.8133	0.0



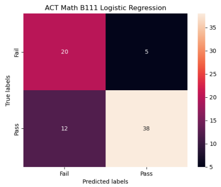
Correlation	P-Value
0.3362	$1.4 * 10^{-35}$



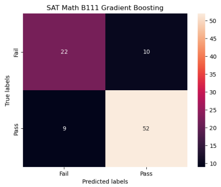


Results

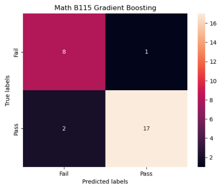
After doing a feature analysis, the following features were determined to have the strongest influence on the model: High School Percentile, High School Readiness Score and High School Score (from US News), SAT and / or ACT score, and the courses taken in high school and their respective subjects.



Accuracy
77.33 %



Accuracy
79.56 %



Accuracy
89.28 %

Limitations and Discussing Results

Limitations: The biggest limitation is incomplete data and the resulting loss of usable entries. Additionally, we were only able to use two class, Math B111 and MathB115, because the other classes did not have enough samples, or the class had to many online courses, because of too different categories of learning environments. Furthermore, we ended up dividing ACT and SAT scores for Math B111, because these tests are different in nature and also vary by year. Moreover, we were not able to build a Z-score. This is because we have a biased dataset, based on the students who enroll at USCB. For Math B115 however, we had to use a conversion chart for the ACT and SAT scores due to limited data [8]. Another limitation is, that we only accounted for students coming-in from public High Schools, which represents the majority of students at USCB. This was a result of the available High School rankings on US News [7].

Discussing Results: Given our limited data and also biased data, we were able to build three well performing models: two models with ACT & SAT split for Math B111, and one model Math B115 with SAT & ACT combined due to fewer samples. While these results are promising, given we are dealing with a dataset involving various different challenges, we want to invest more time in our research in order to achieve even better results, and potentially a model the university can use to classify incoming students.

Future Work

- Implement better pooling algorithm
- Add new data as semesters complete
- Add Z-Score based on year for ACT & SAT
- Build a model to classify both Pass / Fail and the course
- Add newer libraries that enable hardware acceleration

References

[1] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[2] Seaborn: Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021.

[3] Pandas: Data structures for statistical computing in python, McKinney, Proceedings of the 9th Python in Science Conference, Volume 445, 2010.

[4] Numpy: Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2. (Publisher link)

[5] Mathplotlib: J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

[6] Plotly: Author: Plotly Technologies Inc. Title: Collaborative data science Publisher: Plotly Technologies Inc. Place of publication: Montréal, QC Date of publication: 2015 URL: https://plot.ly

[7] US News High School Rankings: https://www.usnews.com/education/best-high-schools/rankings-overview

[8] Bradley University Conversion chart: https://www.bradley.edu/offices/student/asc/faculty-staff/conversion-table/