Using Data Science to Support Student Success

Utilizing Exploratory Data Analysis and Machine Learning to Predict and Summarize College Student Academic Performance

Tim Kaufman, Garrett King

Faculty advisors: Keith Hubbard (Ph.D.) | Dipak Singh (Ph.D.)

Department of Computer Science, Stephen F Austin University

This research was funded in part through Stephen F. Austin State University's *President's Innovation Fund*.

Objective

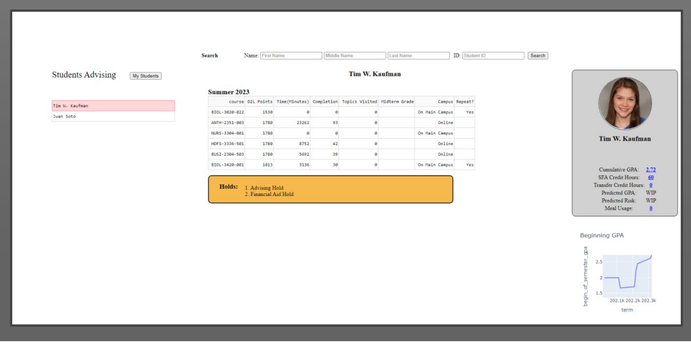
We aim to use machine learning and exploratory data analysis to find unseen patterns that may be of interest to advisors to promote student success.

Week 10 Engagement			Percent Change					Weekly Totals				
			D2L points	D2L time	D2L logins	Meals	AARC	D2L points	D2L time	D2L logins	Meals	AARC
421013	Amelia	Williams	-6	-100	-100			23	0.0	0.0	0.0	0.0
714239	Ethan	Anderson	-1	-71	29			79	0.03	3.0	0.0	0.0
52310	Lily	Thompson		-59	-18			0	0.2	1.0	0.0	0.0
11084	Owen	Mitchell	-4	-86	-27	87		64	0.45	3.0	11.0	0.0
829410	Scarlett	Harris	0	153	-37	41		75	2.44	3.0	11.0	0.0
31048	Noah	Campbell	100		50			100	0.0	1.0	0.0	0.0

Data Cleaning

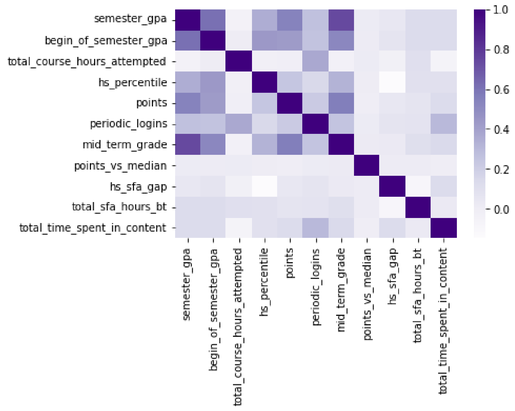
Before processed through a machine learning model, data must first be cleaned. This includes imputing missing values and standardizing features such as date format. Throughout our dataset there were many discrepancies with differing date formats and course codes that had to be rigorously cleaned before model fitting.

semester_gpa	974
begin_of_semester_gpa	6688
total_course_hours_attempted	0
hs_percentile	15210
points	3557
periodic_logins	5132
mid_term_grade	4883
points_vs_median	2830
hs_sfa_gap	704
total_sfa_hours_bt	0
total_time_spent_in_content	171



Data Exploration

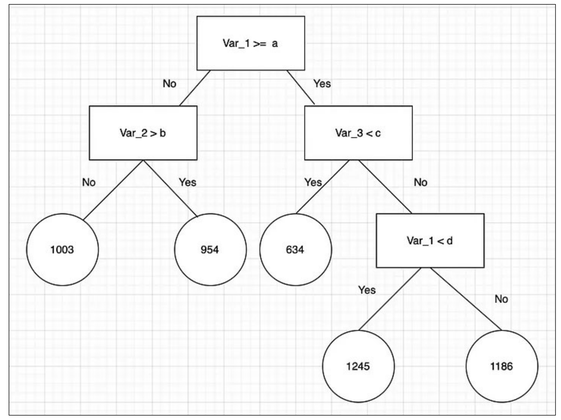
When creating a regression model, a great first step is to gather information about the linear correlation values between your features and target variable. In the case of this project, we found that the highest linearly correlated feature to a student's GPA at the end of a semester, is their midterm grade. This is a great piece of insight, however, the biggest setback is that we do not have data for a student's midterm grade until week 8 of the semester. To overcome this, we use 2 different models; one of the models is used prior to week 8 and the other model is used after midterm grades are finalized.



Random Forest Regressor

Random Forest Regressor has two primary traits to its functions. The first feature is how it samples data; random forest regressor takes random groups, or "forests", of data features to estimate a label value. The second model trait is incorporating decision trees in making choices of how to interpret information.

After averaging the decision trees and calculating the error, the Random Forest Regressor evaluated more accurately than the previous Multivariate models, as is detailed below.



Beheshti, Nima. Towards Data Science, March 2 2022, https://towardsdatascience.com/random-forest-regression-5f605132d19d. November 6 2023

Conclusions

With the current random forest regressor, we have a root-mean-squared error of ~0.633. This means that we are predicting a student's end-of-semester GPA within a margin of on average (+|-) 0.633. We are currently researching and implementing different techniques in order to reduce this error as low as possible

Missing data is the biggest issue with our current model. We expect that when we are able to find better solutions to imputing this missing data, the performance of our model will improve drastically.

Improvements

Feature engineering has proved to be very helpful with improving our model. For example, logging into D2L multiple times within the span of a few minutes does not indicate student engagement. However, only counting 1 login per 10 hours does.

In the future, we would like to complete the advisor dashboard website to make guiding students towards success easier and more robust. On top of this, we are currently working on implementing Time-Series Forecasting models with this data because we believe that a given student's performance is highly dependent on how they have performed in the past weeks or semesters.