



Evaluation of Decision Tree Models to Forecast GPA of Entering Freshmen

Prepared by the Office of Institutional Research, Planning & Effectiveness, June 6, 2017

Because students who earn low first term grade point averages (GPAs) are at significant risk of leaving Stony Brook and not graduating in four years, a data mining system was created to predict first-time full-time freshmen grade point averages (GPAs) prior to entry, three weeks after the start of the term, and six weeks after the start of the term. Principal findings:

- Models demonstrate significant accuracy (actual average GPA for most at-risk groups was within 0.2 grade points of predicted average GPA) and may be used to develop interventions for at-risk students up to 8 weeks into the term and prior to receiving grades
- Lower first term average GPAs were associated with
 - Weaker high school records
 - Lower activity in the BlackBoard learning management system (LMS) among students with weaker high school records
 - Enrollment in courses with higher historical rates of students earning course grades of D, F, and W.
 - Students receiving mid-term reports of weak performance from instructors

Methods

SAS Enterprise Miner was used to create three decision tree models using fall 2014 and fall 2015 data available at three time points—Day 1, the end of week 3, and the end of week 6. A wide range of available data was used including test scores, financial aid data, demographics, advising and tutoring center interactions, and BlackBoard logins. For more detail regarding the modeling methods and outcomes, please see the accompanying paper. A list of the data incorporated into the model can be found in the appendix. Sections of the decision trees for some of the nodes with the lowest predicted GPA's are also presented in the paper.

The IDs of some of the students predicted to have lowest GPA outcomes on Day 1 and at the end of Week 3 were provided to the heads of Academic Advising, EOP, and the Undergraduate Colleges for the first time in fall 2016.

Model Evaluation

The accompanying spreadsheet illustrates the three decision tree models and along with the model predictions and actual fall 2016 GPA's. The decision rules appear in text and the groups suggested for interventions are highlighted in yellow. On inspection one finds the predicted and actual results are quite concordant, with larger divergence most often in nodes with lower frequencies. Since the BlackBoard data were not saved and stored prior to the initial data request for this project in fall 2014, the amount of data relative the number of variables is rather limited. We have noted that some GPA outcomes are higher than expected due to lower D, F, and W rates in some fall 2016 classes. That will be examined in more detail and the model will undergo further training to obtain more finely tuned results. Midterm grade reports were included for the first time in the fall 2016 model. Their predictive power holds promise for work that will commence to identify more BlackBoard metrics to add to the data. For details regarding this project, please see Galambos, N. (2016) "[Mining data to create a system to identify at-risk freshmen](#)" Paper presented at the North East Association for Institutional Research Annual Meeting, Baltimore, Md.

Table 1. Prior to entry model – differences between predicted and actual first term GPA

Model criteria			Predicted avg. GPA	N	Actual avg. GPA	Difference
HS GPA < 86.5	EOP Student = 'Yes'		2.47	31	2.36	-0.11
	EOP Student = 'No'		2.73	141	2.78	0.05
HS GPA >= 86.5 & HS GPA < 90.245	HS Avg. SAT Math Score < 530	DFW Total STEM Courses >= 2	2.30	20	2.57	0.27
		missing(DFW Total STEM Courses) = 1 or DFW Total STEM Courses < 2	2.72	123	2.68	-0.05
	HS Avg. SAT Math Score < 671 AND HS Avg. SAT Math Score >= 530 or MISSING	Math Placement < 1.5	0.67	1	3.25	2.58
		Math Placement < 2.75 AND Math Placement >= 1.5 or MISSING	2.94	53	2.83	-0.11
		Math Placement < 5.5 AND Math Placement >= 2.75	2.77	154	3.00	0.23
Math Placement >= 5.5	3.14	46	2.99	-0.15		
HS Avg. SAT Math Score < 685 AND HS Avg. SAT Math Score >= 671		2.61	0			
HS Avg. SAT Math Score >= 685		3.22	72	3.17	-0.05	
HS GPA < 93.935 AND HS GPA >= 90.245	HS Avg. SAT CR + M Score < 1029.5	Total STEM Units >= 6.5	2.71	140	2.87	0.16
		Total STEM Units < 6.5 or MISSING	3.14	86	2.95	-0.19
	HS Avg. SAT CR + M Score < 1192 AND HS Avg. SAT CR + M Score >= 1029.5 or MISSING		3.01	265	3.05	0.04
	HS Avg. SAT CR + M Score >= 1192 & HS Avg. SAT CR + M Score < 1435		3.35	163	3.34	-0.02
HS Avg. SAT CR + M Score >= 1435		3.82	4	3.59	-0.23	
HS GPA < 96.305 AND HS GPA >= 93.935 or MISSING	HS Avg. SAT CR + M Score < 737.5		3.73	5	2.76	-0.97
	HS Avg. SAT CR + M Score < 1033.5 AND HS Avg. SAT CR + M Score >= 737.5	DFW Total Units Taken < 6.5 or MISSING	3.30	278	3.03	-0.28
		DFW Total Units Taken >= 6.5	2.96	109	3.21	0.24
	HS Avg. SAT CR + M Score < 1243.5 AND HS Avg. SAT CR + M Score >= 1033.5 or MISSING		3.24	430	3.05	-0.19
	HS Avg. SAT CR + M Score >= 1243.5	Math Placement < 3.5 or MISSING	3.68	45	3.61	-0.07
Math Placement < 4.5 AND Math Placement >= 3.5		3.05	4	2.98	-0.07	
Math Placement < 5.5 AND Math Placement >= 4.5		3.38	8	3.07	-0.31	
Math Placement >= 5.5		3.64	39	3.50	-0.14	
HS GPA >= 96.305	SAT 1600 Score < 1175		3.28	48	3.12	-0.17
	SAT 1600 Score < 1195 AND SAT 1600 Score >= 1175		3.04	29	3.26	0.23
	SAT 1600 Score < 1265 AND SAT 1600 Score >= 1195		3.30	111	3.40	0.10
	SAT 1600 Score < 1395 AND SAT 1600 Score >= 1265	DFW Total STEM Units Taken < 4.5	3.58	50	3.40	-0.18
		DFW Total STEM Units Taken < 10.5 AND DFW Total STEM Units Taken >= 9 or MISSING	3.74	153	3.49	-0.25
		DFW Total STEM Units Taken < 9 AND DFW Total STEM Units Taken >= 4.5	3.38	110	3.34	-0.04
DFW Total STEM Units Taken >= 10.5		3.27	12	3.50	0.24	
SAT 1600 Score >= 1395 or MISSING		3.59	138	3.56	-0.03	
Honors College = 'Yes'		3.75	63	3.76	0.01	

Highlighted cells indicate at-risk groups for potential interventions. Gray text represents small cell sizes and differences between predicted and actual GPA should generally be discounted

Table 2. Week 3 model – differences between predicted and actual first term GPA

Model Criteria				Predicted avg. GPA	N	Actual avg. GPA	Difference	
HS GPA >= 93.9	Disbursed Athletic Aid >= 11358.1 or missing	SAT 1600 < 1255		3.16	413	3.23	0.07	
		SAT1600 >= 1255 or missing		3.46	947	3.41	-0.06	
	Disbursed Athletics Aid < 11358.1	Math Placement >= 6.5	BB Total Logins: wk 3 < 26		2.39			
			BB Total Logins: wk 3 >= 26 or missing	SAT Math < 655	3.28			
				SAT Math >= 655 or missing	3.61			
		Math Placement < 6.5 or missing	HS GPA < 96.165	DFW Total STEM Units < 7	3.30	4	3.27	-0.04
				DFW Total STEM Units >= 7 or missing	2.94	0		
			HS GPA >= 96.165	BB Total Non-STEM Logins: wk 3 < 38 or missing	3.30	1	2.92	-0.38
			BB Total Non-STEM Logins: wk 3 >= 38	3.58	3	3.50	-0.09	
	HS GPA < 93.9 or missing	Total BB Logins: wk 3 < 61	BB Logins Per Crs: wk 3 < 2.083		0.95	1	1.25	0.30
BB Logins Per Crs.: wks3 >= 2.083 or missing			DFW STEM Total STEM Units < 5 or missing	2.72	203	2.55	-0.17	
			DFW Total STEM Units >= 5	2.27	24	2.71	0.44	
HS Avg. SAT CR + M + W Score >= 1571.5 or missing			Total BB Logins Per Crs.: wk 3 < 5.23		2.34	40	2.14	-0.20
			Total BB Logins Per Crs.: wk 3 >= 5.23 or missing		2.87	279	2.79	-0.09
Total BB Logins: wk 3 >= 61 or missing			Total AP STEM Credits < 4		3.14	36	3.45	0.31
		HS Avg. SAT CR Score >= 565	Total AP STEM Credits >= 4 or missing	3.22	126	3.33	0.11	
			SAT Writing < 655 or missing	3.55	50	3.38	-0.17	
			SAT Writing >= 655	2.95	675	3.01	0.06	
		HS Avg. SAT CR Score < 565 or missing	Disbursed Athletic Aid < 33989.1 or missing	Disbursed Grant Aid < 12861 or missing	2.66	120	2.93	0.27
			Disbursed Grant Aid >= 12861	3.03	9	3.45	0.42	
	Disbursed Athletic Aid >= 33989.1	Total AP Credits < 11 or missing	3.00	0				
		Total AP Credits >= 11						

Highlighted cells indicate at-risk groups for potential interventions. Gray text represents small cell sizes and differences between predicted and actual GPA should generally be discounted

Table 3. Week 6 model – differences between predicted and actual first term GPA

Model Criteria				Predicted avg. GPA	N	Actual avg. GPA	Difference	
Classes Listed on Mid-term Grade Report >= 1				2.31	389	2.42	0.11	
Classes Listed on Mid-term Grade Report = 0 or missing	HS GPA >= 94.08	Math Placement >= 6.5		3.56	188	3.54	-0.02	
		Math Placement < 6.5 or missing	Total BB Non-STEM Logins: wk 6 >= 31		3.46	232	3.57	0.11
			Total BB Non STEM Logins: wk 6 < 31 or missing	Total AP STEM Courses < 2 or missing	3.09	309	3.28	0.19
				Total AP STEM Courses >= 2	3.38	312	3.46	0.08
	HS GPA < 94.08 or missing	BB Logins Per Course: wk 3 >= 7.82 or missing	HS Avg. SAT CR Score < 567.5 or missing		2.97	996	3.08	0.11
			HS Avg. SAT CR Score >= 567.5	Total AP STEM Credits < 7 or missing	3.21	184	3.34	0.14
				Total AP STEM Credits >= 7	3.45	111	3.46	0.01
		BB Logins Per Course: wk 3 < 7.82	Total BB Logins: wk 6 < 32	DFW STEM Units < 5 or missing	2.62	0		
				DFW STEM Units >= 5	1.78	0		
			Total BB Logins: wk 6 >= 32 or missing	HS Avg. SAT CR + M + W Score < 1601.5 or missing	2.64	142	2.58	-0.06
		HS Avg. SAT CR + M + W Score >= 1601.5	2.96	68	3.01	0.05		

Highlighted cells indicate at-risk groups for potential interventions. Gray text represents small cell sizes and differences between predicted and actual GPA should generally be discounted