Identifying At-Risk Students for Targeted Academic Support

Paul Horton

December 2020

1 Problem Statement

Universities have complex requirements for admission decisions. Admissions teams must consider maintaining an intelligent, diverse student body while constrained by budget requirements and other quotas for sports teams, bands, and alumni relations.

Universities will thus admit students with a range of backgrounds, previous academic achievements, and preparedness. Many of these students will be ready for college while others will struggle. All colleges have resources available to help struggling students. Just at Georgia Tech, the Registrar's Office lists academic advisement, counseling, and tutoring as available resources. [2] The academic support teams cannot provide all these services to each student. Nor can they waste resources by contacting or appealing to students who do not need the help. Instead, they need to be efficient.

It is in the interest of the university to effectively target unprepared students so that they can get the resources they need to be successful. A successful student will benefit the university in the long term by improving the image, donating, and strengthening the alumni network. Unprepared students need to be identified as early as possible. The further along they get without receiving help, the bigger the hole they will be in and the lower their confidence. It will be most effective to target them prior to beginning their first year.

Algorithms involved in admissions processes are very controversial as Amazon showed after it stopped using its biased recruiting algorithm. [1] This issue is mostly avoided in this case by being supportive and targeting those in need. Although, it still needs to be easily explained. This tool could potentially be used by public institutions where it will be questions and held accountable for decisions that it recommends.

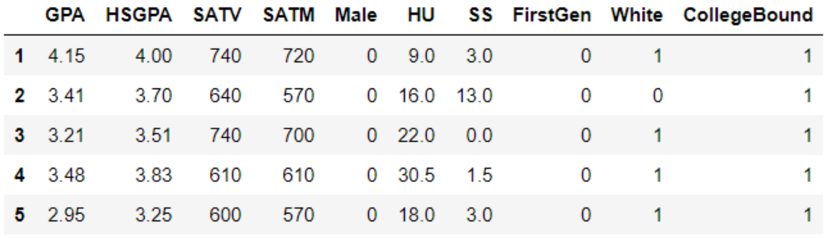
2 Data

The data come from a sample of 219 first year students at a Midwestern college. This is a clean data set, and I will not have to do significant pre-processing.

There are 10 variables:

GPA – First year grade point average on a 4.0 scale.
HSGPA – High School grade point average on a 4.0 scale.
SATV – SAT Verbal score out of 800.
SATM – SAT Math score out of 800.
Male – Gender where a value of 1 indicates male and 0 for female.
HU – Credit hours of humanities courses taken in high school.
SS – Credit hours of social science courses taken in high school.
FirstGen – An indicator if this student is attending college as the first in their family.
White – Race where a value of 1 indicates white and 0 for other.
CollegeBound – Student comes from high school where over 50 percent attend college and 0 otherwise.

GPA is the variable that we are trying to predict and is thus the dependent variable. GPA, SAT, and credit hours are technically discrete, although, I will treat them as continuous for this analysis. The other variables are binary. We do not have any variables which indicate the department or major that the students have selected. Variables have a wide range of values so it will be important to scale these prior to using in a KNN model.



	GPA	HSGPA	SATV	SATM	Male	HU	SS	FirstGen	White	CollegeBound
1	4.15	4.00	740	720	0	9.0	3.0	0	1	1
2	3.41	3.70	640	570	0	16.0	13.0	0	0	1
3	3.21	3.51	740	700	0	22.0	0.0	0	1	1
4	3.48	3.83	610	610	0	30.5	1.5	0	1	1
5	2.95	3.25	600	570	0	18.0	3.0	0	1	1

Figure 1: Sample of data table.

3 Methodology

The objective is to deliver a model which will accurately predict which students will need additional academic resources to succeed in their first year. This first requires defining what an unsuccessful student is. There are two approaches you could use for this:

1) Set a GPA target and label everyone who is below that threshold as someone who needs additional support.
2) Use a bottom percentile of students to define who needs help.

The advantage to using the first option is that teachers are likely already grading on a scale. They may only give 10 percent of students a D or below. The disadvantage is that there is a phenomenon known as grade inflation where average GPAs continue to rise.[3] The threshold would need to be reevaluated every 5 years to ensure the right students are being targeted.

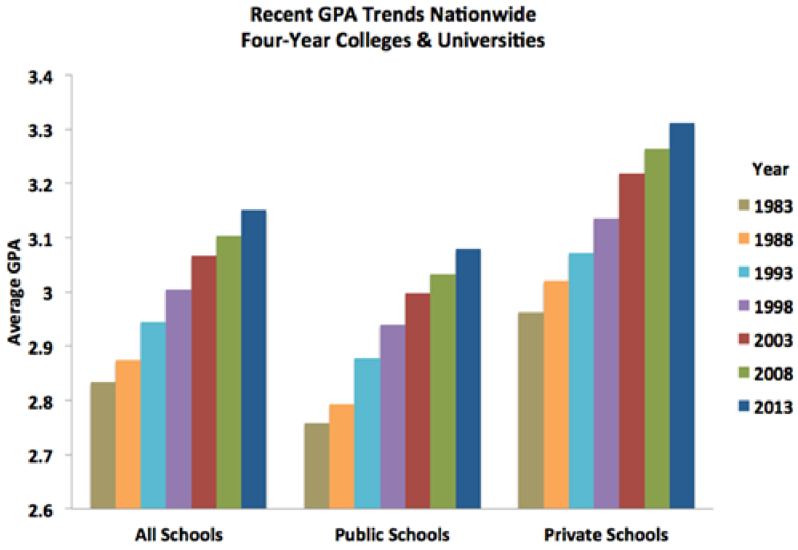


Figure 2: Grade inflation over last four decades.

The second method still requires calculating a cutoff GPA, but it would reflect any changes in how teachers are grading. This would require using the previous year's results for determining a cutoff. Both methods are reasonable, but I will use the second. I will assume that any student in the lowest quintile is one who needs additional support. I plotted the cumulative density estimation for first year GPA in Figure 3 below and determined that 2.7 will be an appropriate GPA threshold. Ideally, I would have information about which students graduated and successfully found a job upon graduation. Long term metrics like those would be more indicative of a successful education rather than only the short term GPA. Those data are more difficult to gather but it would support the definition of what is a successful student.

Because one of the objectives is to have an explainable model, I will not focus on neural networks and random forest models. Instead, I will prioritize classification models which have results that are easier to interpret and explain.

The types of classification errors do not have the same significance in weight. It would be costly for a struggling student to not receive information about services. In contrast, it would be less significant for marginal or successful student to receive information about these services. I will weight the former error to be five times more costly than the latter when evaluating different models.

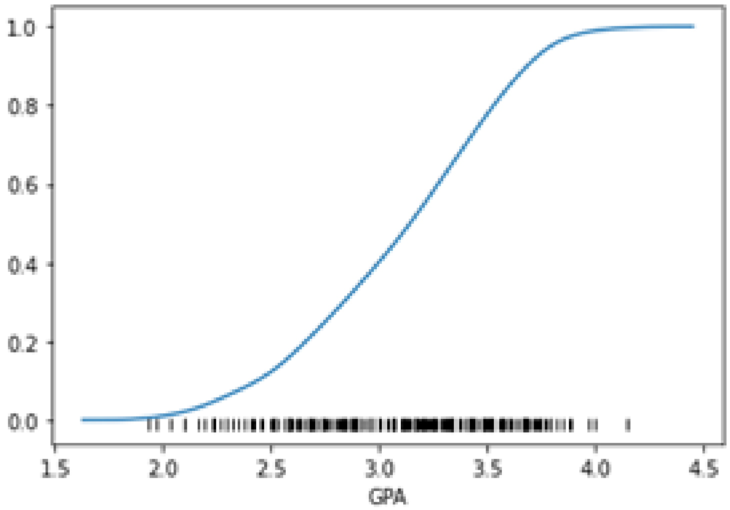


Figure 3: Cumulative density function for first year GPA.

4 Analysis

One aspect of this problem that will complicate classification is that this student sample will already have gone through one round of screening during the application process. The admitted students will have a level of achievement and similarity that makes it difficult to predict who will be successful. As you can see in the plot of the first two principal components below, there is not a clear boundary or group for these two groups of students. It is difficult to imagine that a linear boundary will perform well classifying within this sample.

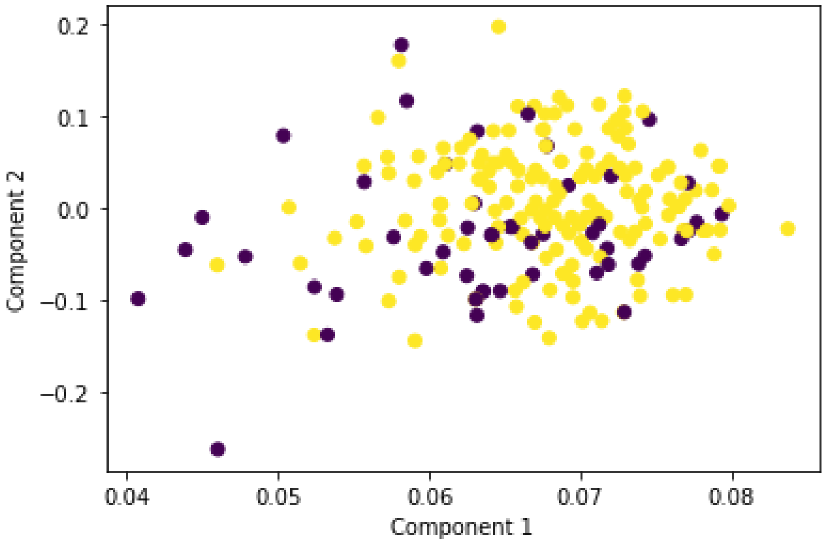


Figure 4: Graph of first two principal components with student success indicator.

One objective of this analysis was to have a transparent model that is easily understandable. To do that, I performed a lasso regression to identify the variables that are most significant to student success. A model with less factors is less likely to over fit the model and it will also be easier to understand. The two factors which were excluded, White and Male, should not be included in the model anyway due to their controversial nature in educational performance studies. After splitting the data into train and test sets, I had it in a format to start building models.

The heat map in Figure 5 shows correlations between variables to help us understand the data. It is not surprising to see that being a first generation college student is negatively correlated with university success. Parents can provide wisdom and share their experience to help prepare their children for what college will be like. It is surprising to see a negative correlation between college bound high schools and college success.

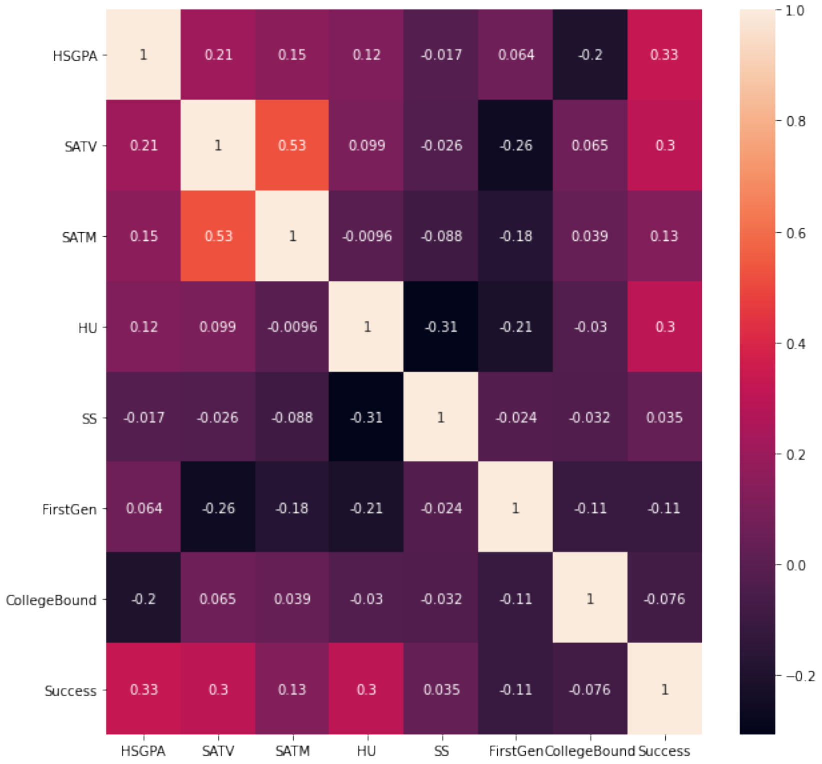


Figure 5: Heatmap of correlation between factors.

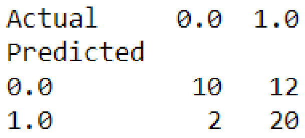
Since I structured this as a classification problem, the models that I initially used were logistic regression, decision tree, and KNN. These three models meet the requirements for transparency and can be tuned for performance. I felt it was also pertinent to have less explicable models such as a neural network, random forest, and Adaboost to determine what level of accuracy you sacrifice for simplicity.

5 Evaluation and Results

Each model will be evaluated on the total cost of miss-classification and the false negative rate. The objective was to identify at-risk students using a false negative cost of 5 times that of false positives.

I started with a logistic regression model because the ability to tune the threshold to bias the classifications would be useful for optimization. The logistic regression function in sklearn has a class weight argument which can be used to shift the classifications. As mentioned in the problem statement, I used a weight of 5 for the bias the error towards false positives and away from false negatives. The false negative rate for this model is 0.17 with a total cost of 22.

Prior to modifying the class weight, the false negative rate for this model was 0.67 and the cost was 41.



```
Actual      0.0  1.0
Predicted
0.0          10   12
1.0           2   20
```

Figure 6: Confusion matrix for weighted logistic regression.

KNN was the next algorithm I tested. The challenge with using KNN in this context is that the outcomes are not evenly distributed. This means that whenever you increase the number of nearest neighbors, you will bias your results against classifying as the minority outcome. The trade off in having a low number of nearest neighbors is the risk of over fitting. To understand the impact of different number of nearest neighbors, I iterated though 2 to 9 to see the difference in performance. The best performing KNN model on the test set had a cost of 29 and a false negative rate of 0.42. Since this is not as good as other models, I will not consider it further.

Decision trees are easily understandable since they can be visualized, as seen below in Figure 7. Unfortunately, they do not have parameters which can be used to bias the classification to improve our model performance. With an unconstrained max depth and number of branches, our best model had a cost of 38 with a false negative rate of 0.5.

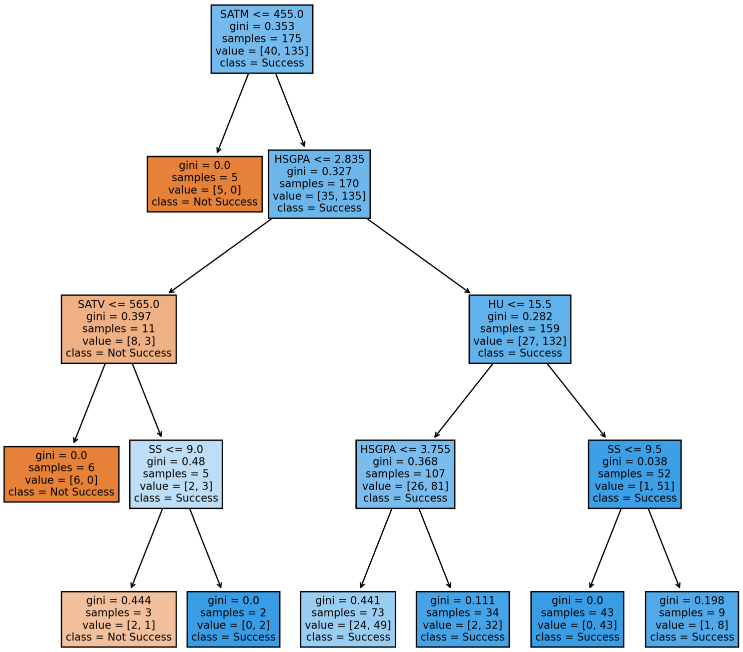


Figure 7: Decision tree with max depth of 4.

I expected improved performance for the neural network, Adaboost, and random forest models. For just accuracy, these models were among the best at rates approaching 80 percent. Unfortunately, they do not have the performance for false negative rate and cost that make them better for this application. The costs for NN, Adaboost, and random forest are 44, 37 and 41 respectively. Their difficulty in explaining also compounds their poor fit.

6 Conclusion

The logistic regression model performed best on this data set but it still did not offer the benefit anticipated. By tuning the class weight, we could correctly identify all at risk students but the model began to classify essentially all of the students as at risk. The school system instead could send notifications and outreach to all students rather than implementing a model like this.

Despite having relevant academic data for incoming freshman, it is clear that there are other factors that are significant in determining whether a student needs help. There is the potential to expand this analysis as schools are increasingly incorporating online learning into the curriculum. Real-time performance for students on assignments would provide a more complete view of the risk of failure. Other models for student performance prediction indicate the need for a continuous process. [4] Schools who want to incorporate a prediction model will therefore need to continue to collect data as a student progresses through the program.

In addition, I recommend collecting information about the major or program that students have selected. Some programs are more difficult than others which means students may not have the same chance of success in each program. A student's SAT score for math could be more a more significant predictor for an engineering major rather than liberal arts.

Additional observations would also improve the robustness of our analysis. There were only 219 observations in the whole data set which left only about 45 for the test set. We further narrowed the scope to false negatives and cost of errors so small, random variations may make models appear better than others.

References

[1] Amazon scrapped 'sexist ai' tool. URL `https://www.bbc.com/news/technology-45809919`.

[2] Georgia institute of technology registrar - academic resources. URL `https://registrar.gatech.edu/current-students/academic-resources`.

[3] Stuart Rojstaczer. National trends in grade inflation, american colleges and universities.

[4] J. Xu, K. H. Moon, and M. van der Schaar. A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, 11(5):742–753, 2017. doi: 10.1109/JSTSP.2017.2692560.