ANALYSIS OF WHITTIER COLLEGE ADMISSIONS DATA: EXAMINING MATRICULATION TO SOPHOMORE YEAR

Jordan Hanson

November 15th, 2018

Whittier College Department of Physics and Astronomy

OUTLINE

1. **Motivation**: *Search admissions data for outliers*
2. **Methods**: *Derive normalized variables from raw data*
3. **Results**: *We can identify outliers*
4. **Next steps**: *Aid gap analysis, classification with more data*
5. **Conclusion**

MOTIVATION

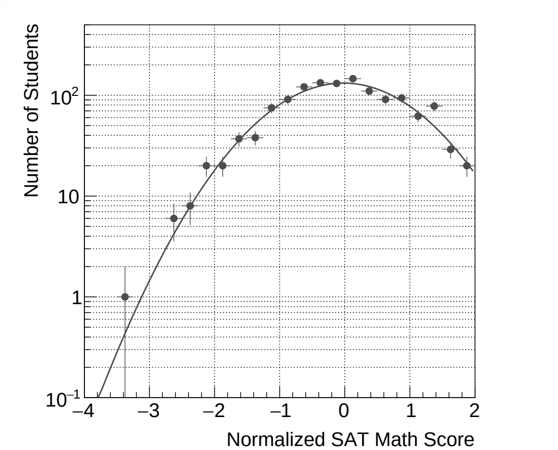

Figure 1: The normalized SAT Math for students in this analysis. The gray line is a normal distribution fit.

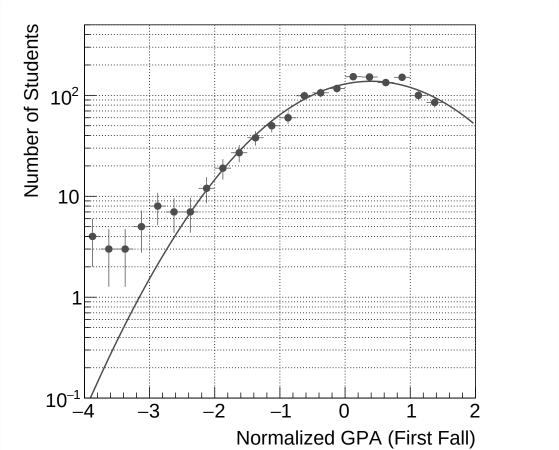

Figure 2: The normalized GPA for the first fall of enrollment for students in this analysis. The gray line is a normal distribution fit.

Questions we look to answer:

- Which students are being admitted with such low scores? How are those scores related?
- Do these students matriculate to sophmore year?

METHODS

1. **Selection**: *From raw data:*
    - Require three SAT scores, and high school GPA
    - Require GPA from Whittier, freshman fall and spring
    - Require INTD 100 grade
    - Require aid gap knowledge
2. **Methods**: *Derive normalized variables from raw data*
    - (See next slide).
3. **Analysis**: *Plot distributions and correlations*
4. **Analysis**: *Fisher discriminants and machine learning*

Normalizing data: Let s_i represent a score (like GPA) for student i, and let $\bar{s}$ and σ_s represent the mean score and standard deviation for the students, respectively. The normalized score is

$$s_{i,n} = \left(\frac{s_i - \bar{s}}{\sigma_s}\right) \tag{1}$$

Criteria	N
Raw data	3119
Require parameters	1346

Table 1: The numbers of students in the analysis. The student data corresponds to 2012-18.

Parameter	Mean	Std. dev.
SAT CR	525.6	82.1
SAT Math	526.5	82.5
SAT Writing	523.0	77.7
GPA HS	3.52	0.42
GPA FF	3.00	0.72
GPA FS	3.01	0.67
INTD100	3.20	0.89
Aid Gap	**539.2**	**14560**

Table 2: The numbers of students in the analysis. The student data corresponds to 2012-18.

ANALYSIS: PLOT DISTRIBUTIONS AND CORRELATIONS

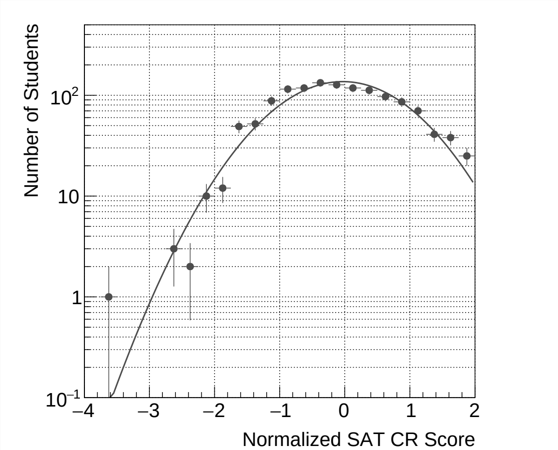

Figure 3: The normalized SAT CR score for all students.

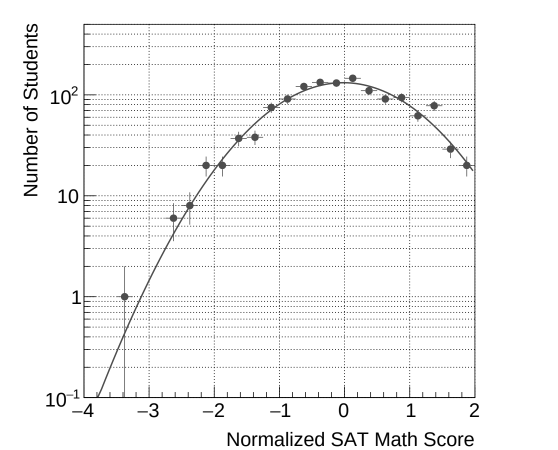

Figure 4: The normalized SAT MATH score for all students.

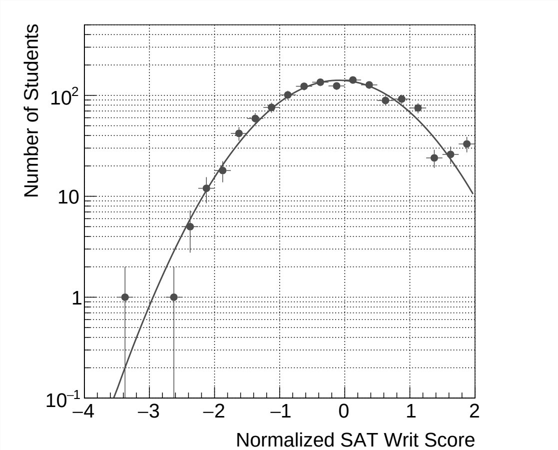

Figure 5: The normalized SAT WRIT score for all students.

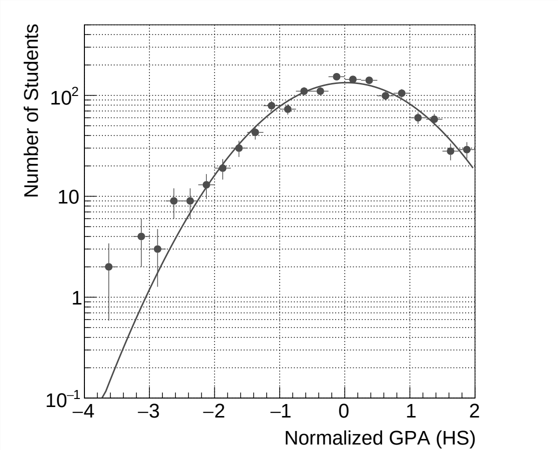

Figure 6: The normalized GPA HS score for all students.

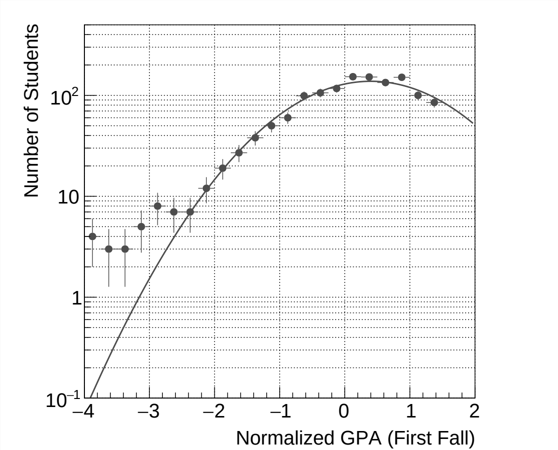

Figure 7: The normalized GPA FF score for all students.

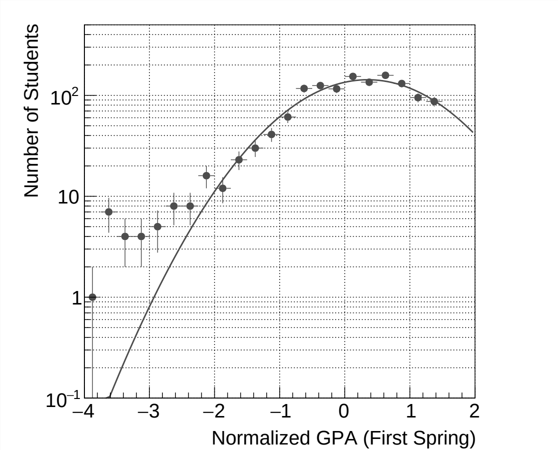

Figure 8: The normalized GPA FS score for all students.

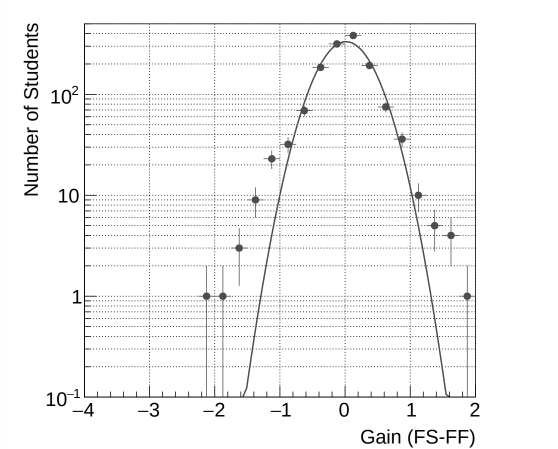

Figure 9: The gain score for all students.

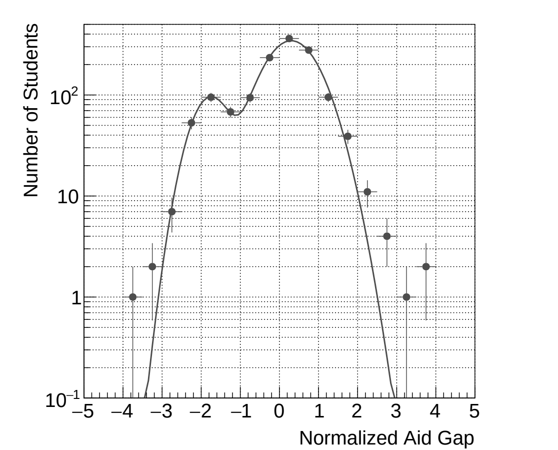

Figure 10: The normalized aid gap for all students. Recall that the standard deviation for the whole set is $14k, and the mean is $500.

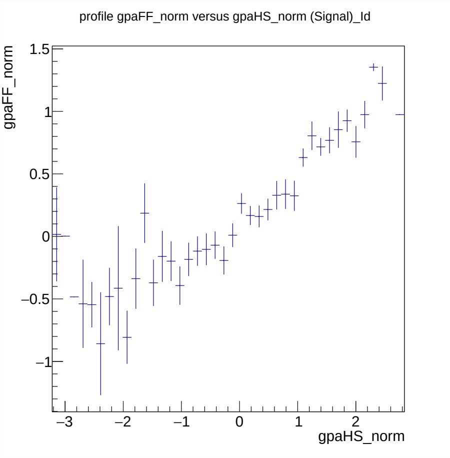

Figure 11: Correlation of GPA FF with GPA HS (do matriculate to sophomore year).

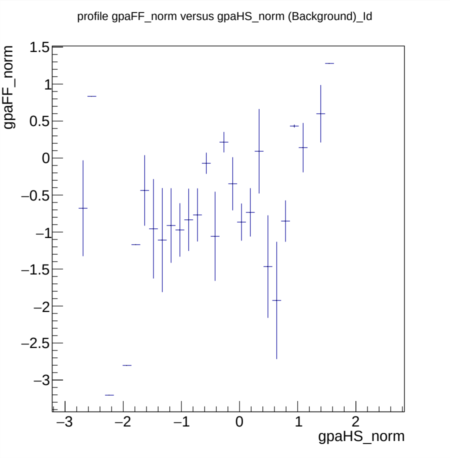

Figure 12: Correlation of GPA FF with GPA HS (do not matriculate to sophomore year).

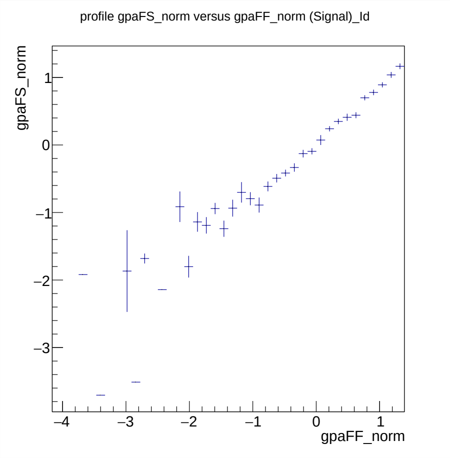

Figure 13: Correlation of GPA FS with GPA FF (do matriculate to sophomore year).

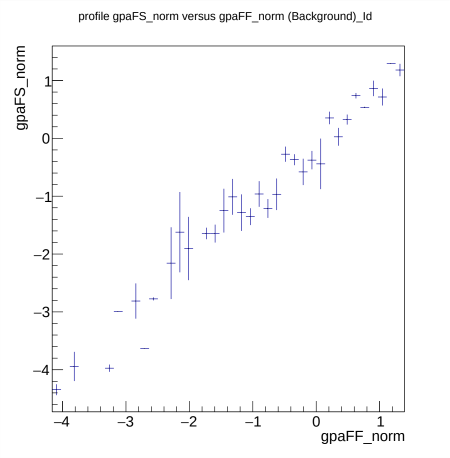

Figure 14: Correlation of GPA FS with GPA FF (do not matriculate to sophomore year).

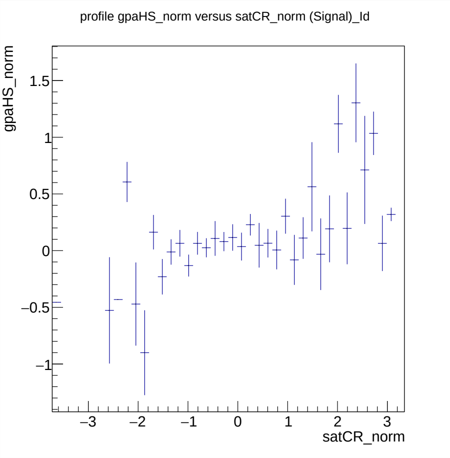

Figure 15: Correlation of GPA HS with SAT CR (do matriculate to sophomore year).

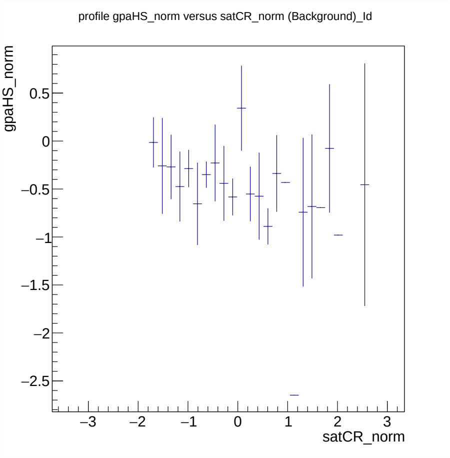

Figure 16: Correlation of GPA HS with SAT CR (do not matriculate to sophomore year).

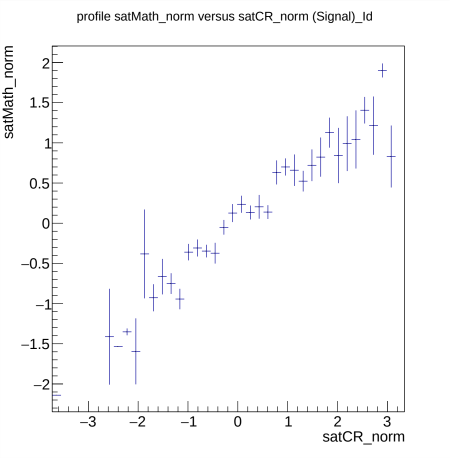

Figure 17: Correlation of SAT MATH with SAT CR (do matriculate to sophomore year).

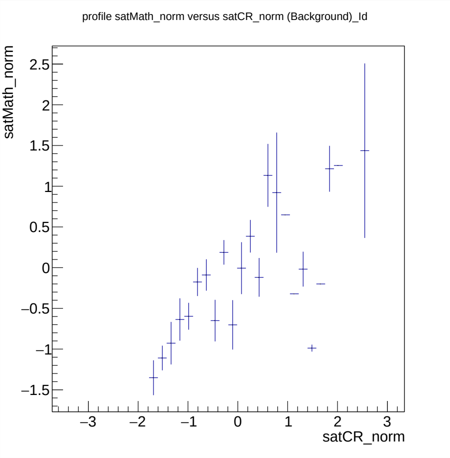

Figure 18: Correlation of SAT MATH with SAT CR (do not matriculate to sophomore year).

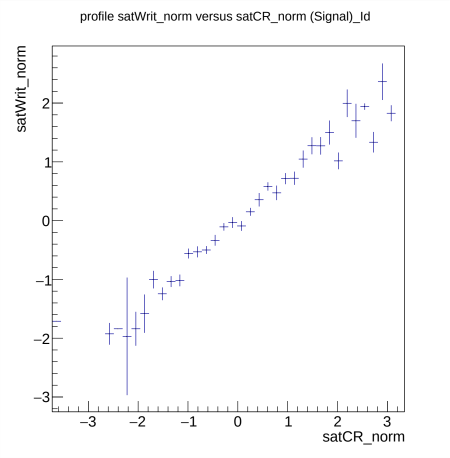

Figure 19: Correlation of SAT WRIT with SAT CR (do matriculate to sophomore year).

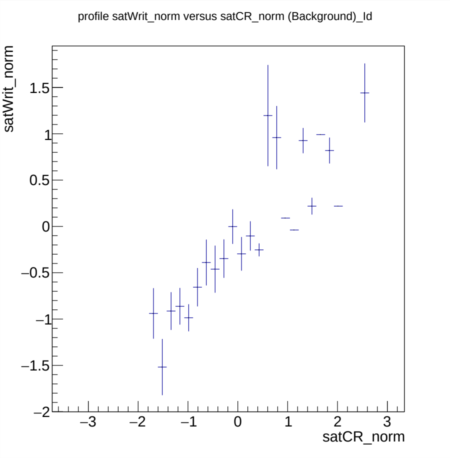

Figure 20: Correlation of SAT WRIT with SAT CR (do not matriculate to sophomore year).

FISHER DISCRIMINANTS AND MACHINE LEARNING

1. A Fisher discriminant is a analysis tool built to distinguish classes. Data point selection is performed by combining features to form *linearly independent* variables and distinguishing them based on their means and standard deviations.
2. k-Nearest Neighbors is *machine learning* technique that classifies data.

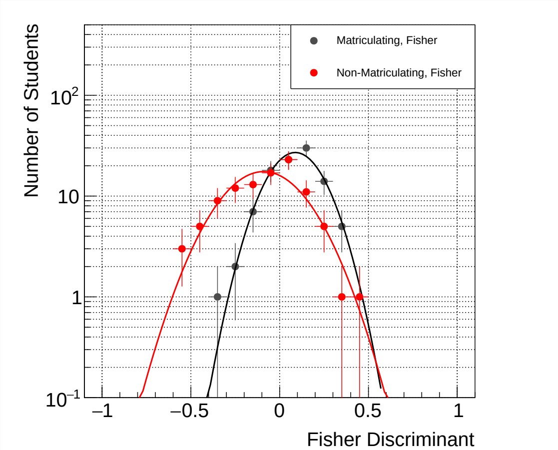

Figure 21: Fisher results. Matriculating: 1043 randomly selected students to train, and 100 to test. Non-matriculating: 104 randomly selected students to train, and 100 to test.

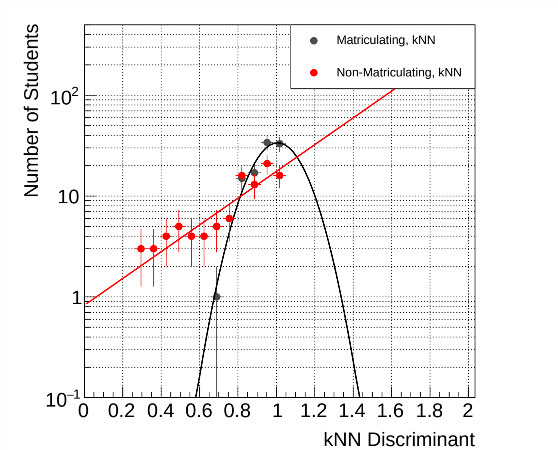

Figure 22: kNN results. Matriculating: 1043 randomly selected students to train, and 100 to test. Non-matriculating: 104 randomly selected students to train, and 100 to test.

NEXT STEPS

1. The aid gap data needs a more thorough analysis.
2. The algorithms presented can identify outliers in the data that (statistically) will not make it to sophomore year. This could be for *any reason*, however.
3. Usually, Fisher and other approaches fully distinguish the distributions:
    - The data is non-linear.
    - We need more data (just the variables indicated, potentially fewer).

CONCLUSION

1. **Motivation**: *We are searching admissions data for outliers.*
2. **Methods**: *We have derived normalized variables.*
3. **Results**: *We have identified outliers corresponding to admitted students with little chance of graduating.*
4. **Next steps**: *The aid gap needs more analysis, and classification will improve with more data*