Predicting College Admissions
Jake Mayer¹, Marc Pittinsky¹, and Yihang Yan¹

¹Department of Computer Science, University of North Carolina at Chapel Hill, NC 27599, USA

Summary: Apparent randomness in the admissions process for major American universities is one of the largest problems facing students who wish to apply. Providing accurate predictions of outcomes for prospective applicants can simplify their application process. We tested numerous models, including a multi-step pre-clustering method, on a dataset of self-reported student profiles and outcomes. Prediction accuracies of up to 85% were achieved with a vanilla MLP.

Motivation
- Millions of students apply to universities on a yearly basis [1]
- There is still randomness in the decision processes. Utilizing GPA and SAT/ACT scores can still have significant outliers in admission statistics
- We wish to investigate the black box of college admissions in an attempt to remove some of the randomness
- A more decisive prediction of one’s admission can allow them to efficiently allocate time and resources into applications for that respective school

Introduction
- Previous approaches have typically focused on predicting graduate program admissions through school-specific models [3]
- We combine features of both the applicant and school in our input to achieve a more versatile and generalized model
- We then test a multi-step learning technique in which we cluster our data based on university or student-specific metrics and train each cluster individually
- By clustering student profiles, we examine our hypothesis that factors used in an admissions decision are largely based on the general performance level of the student
- By clustering universities, we examine our hypothesis that schools with similar attributes utilize similar factors and weights in making admissions decisions
- By comparing our results from differing models and then applying them to a baseline truth of university admission rates, we examine the success of our procedure

Method
- Baseline results are established using a ridge classifier and logistic regression on a holdout data set
- As a preliminary approach, we construct a simple ANN with 3 hidden layers of 21 nodes each (one node per feature)
- Using K-means, we then cluster universities on their acceptance rate, average GPA, and average test score
- A logistic regression and MLP are trained for each cluster
- We take the same approach as above, instead clustering students on GPA, class rank, and test score
- 10-fold validation is used to obtain performance metrics

Data
- Roughly 9,000 self-reported undergraduate student profiles and application outcomes for specific universities after balancing
- University specific undergraduate admission statistics for each university present in student profiles
- SAT and ACT Scores were normalized using Equation 1. Class rank and GPA were normalized through simple operations
- Notable features include acceptance rate, average GPA, student class rank, student GPA, and student test scores

\[
\text{TextScore} = \frac{\text{ACT or 3(average SAT(SATscores))}}{36}
\]

Equation 1: The method for normalizing standardized testing in student profiles

Experimental Results
- Vanilla MLP performed best with roughly 85% accuracy while the clustering approach yielded roughly 83%
- Utilizing the best performing vanilla MLP, experimental admission rates were fairly consistent with university-reported admission rates in universities with large amounts of samples

Acceptance Rate Accuracies

<table>
<thead>
<tr>
<th>Cluster</th>
<th>GPA</th>
<th>Class Rank</th>
<th>Test Score</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.197</td>
<td>0.326</td>
<td>0.675</td>
<td>77%</td>
</tr>
<tr>
<td>2</td>
<td>3.781</td>
<td>0.136</td>
<td>0.789</td>
<td>89%</td>
</tr>
<tr>
<td>3</td>
<td>2.944</td>
<td>0.049</td>
<td>0.902</td>
<td>67%</td>
</tr>
</tbody>
</table>

Table 2: The average precisions normalized by support predicting acceptance and rejection separated into 3 clusters by concrete grade metrics

Conclusions
- We suggest that there is an upper bound in predicting college admissions attributed to factors that are more difficult to quantify such as essay and extracurriculars
- Model performance was much higher when trained on higher performing students
- This then suggests that underperforming students may rely more on metrics such as essay and extracurriculars
- Universities with similar testing metrics have unique admissions processes

References:
[2] CollegeData.com, Admissions Tracker