Predicting College Admissions
COMP 562 - Introduction to Machine Learning

Github: https://github.com/peakay/CollegePredictions

Jake Mayer
Marc Pittinsky
Yihang Yan

Abstract

In this paper, we present our results of the application of machine learning on profile-based prediction of outcomes in college admissions. We describe previous work on similar or related problems and note the aspects which we plan to adopt or modify. We go over challenges regarding how the data set was compiled, and then normalized to maximize acceptable sample count. We then present our baseline model, followed by discussion of our novel multi-task clustering approach. We show that our baseline models resulted in an accuracy of 85% when predicting the outcome of a student, while our novel approach performed slightly worse with an accuracy of about 83%. We analyze and discuss possible shortcomings of these models and propose avenues for future research.

1. Introduction

Millions of students apply to universities on a yearly basis (3). College Admission decisions are a black box to students despite large amounts of admission data being publicly available. Simply utilizing GPA and SAT/ACT scores has significant outliers in admission statistics. Students with the same credentials may be accepted one year while denied the next. In order to better understand admission trends, a multi-step model was implemented based on the hypothesis that similar universities make similar admission decisions. Similarly, the model was applied to students by clustering on their respective metrics.

2. Literature Review

2.1. Previous Approaches

Existing applications of machine learning in this field have mainly concerned results in graduate program admissions, and no formal research was found on previous applications to undergraduate admissions. However, many of the foundational principles remain the same and are worth noting. We discuss two such studies in detail below.

We first examine the approach in (2), which utilizes a Naive Bayes classifier to derive the likelihood of a given student's admission to a various graduate programs. Data is taken from the "MS-in-US" Facebook community, and provides insight into student profile features such as work experience, GRE score, TOEFL score, undergrad university and major. The data set is divided into subsets by school, each of which is used to train a separate model corresponding to that school’s program. New student information is then evaluated against each of these models, producing a probability of admission to each respective program.

The research study in (4) takes a similar approach, instead utilizing the results to provide guidance to the graduate schools in assessing potential candidates. As a product of the University of Texas at Austin’s School of Computer Science, the algorithm is provided with full application materials from students seeking entry into the PhD program. From this, an ideal subset of features is selected and a model is developed using a logistic regression classifier. Results for each candidate are submitted to the admissions board for further consideration.

2.2. Adaptations and Changes

Our model differs drastically from previous work in this domain on account of several factors. To our knowledge, it is the first attempt of utilizing a neural network to address the problem posed. We are able to train this efficiently using a volume of data much larger than those of previous studies. While many of the student profile features resemble those discussed above, we opt to include a corresponding
feature set for each school. An instance of our data is then the features of a student and of the school to which they are applying. Using this, we can train profiles from multiple schools under the same model, which we suggest will increase the accuracy and generalization of our approach.

3. Data Set

Data was obtained in the form of self-reported profiles from CollegeData.com. CollegeData has recorded approximately 1000 admission profiles per university detailing the applicants high school, GPA, SAT, ACT, Class Rank, and athlete status between 2009 and 2020 (1). There were approximately 1,300 Universities total in the data set. Approximately 9,000 rows were used in training and testing the model after cleaning. In order to decrease the loss of data in cleaning, standardized test scores were consolidated into one metric called test score ($TS$). The method for producing this metric is described in Formula 1. Student class rank and GPA were normalized using simple operations. We note that we lose selectivity based on SAT test scores per subject, and preference of a standardized test of a given school. However, this method allows us to retain an additional 2000 samples.

$$TS = \frac{ACT \text{ or } SAT_{to\text{-}ACT}[3 \text{ avg}(SAT_{sub})]}{36}$$  \hspace{1cm} (1)

4. Baseline Approach

4.1. Model

Our baseline approach consisted of three different learning models. A ridge classifier, logistic regression, and multilayer perceptron (MLP) were used to classify whether student profiles were accepted or denied. These models were trained on our combined university-student data set. The MLP, shown in Figure 1, consisted of 3 hidden layers with 8 nodes each, utilizing ReLU for activation and Adam for weight optimization. 10-fold cross validation was applied to each model over 100 iterations and the performance was averaged.

5. Novel Approach and Commentary

5.1. Model

A multi-task model was implemented. Utilizing k-means, universities were clustered by their average test score, average acceptance rate, and average GPA into 10 clusters. A multilayer perceptron was then trained on the data split by the corresponding clusters. The same approach was applied to students with 3 clusters on the metrics of unweighted GPA, test score, and class rank. 10-fold validation was then used to evaluate performance and the weighted precisions by support were averaged to describe an overall performance. The visualized model is present in Figure 2.

![Figure 2. Diagram demonstrating the structure of the multi-task model applied to students and universities separately.](image)

5.2. Commentary

The intuition behind the multi-task model is based on the hypothesis that universities with similar metrics will have comparable admission processes and students with similar metrics will have comparable admission rates. Therefore, clustering before training is an attempt to isolate these trends in the data set. Additionally, clustering has the added benefit of improved generalization, such that the classes are wider (and thus have more samples per class).

6. Results

6.1. Baseline Results

The three baseline models performed comparably within a range of around 2%. Ridge regression was the least effective model while the MLP proved to predict with the best accuracy. The precisions, F1 scores, and recall are detailed in Table 1. The MLP achieved an accuracy of approxi-
Figure 3. The precisions, recalls, and F1 scores for each of the models trained on the data set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ridge Classifier</td>
<td>0.833</td>
<td>0.830</td>
<td>0.829</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.840</td>
<td>0.838</td>
<td>0.838</td>
</tr>
<tr>
<td>Multilayer Perceptron</td>
<td>0.849</td>
<td>0.849</td>
<td>0.849</td>
</tr>
</tbody>
</table>

Table 1. Performance metrics for each baseline models, as shown in Figure 3.

6.2. Novel Results

K-means clustering was fine tuned on both the university and student data through binary search. A cluster size of 10 for the universities and 3 for the students were found to provide the maximum performance. Clustering on student metrics had better performance than clustering on university metrics. The average results of both methods over 100 iterations are shown in Table 2 below.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student Clustering</td>
<td>0.829</td>
<td>0.823</td>
<td>0.823</td>
</tr>
<tr>
<td>University Clustering</td>
<td>0.765</td>
<td>0.762</td>
<td>0.762</td>
</tr>
</tbody>
</table>

Table 2. Comparison of performance metrics for the two clustering methods attempted.

The bad performance of the clustering approach could be due to several factors. One of the main factors seems to be the isolation of poor performances in under-performing students. Table 3 demonstrates the difference in performance by cluster with the corresponding student cluster means.

<table>
<thead>
<tr>
<th>Support GPA</th>
<th>Class Rank</th>
<th>ACT Score</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>327</td>
<td>2.83</td>
<td>0.46</td>
<td>23.4</td>
</tr>
<tr>
<td>845</td>
<td>3.46</td>
<td>0.25</td>
<td>26.3</td>
</tr>
<tr>
<td>1451</td>
<td>3.88</td>
<td>0.08</td>
<td>29.9</td>
</tr>
<tr>
<td>2623</td>
<td>3.61</td>
<td>0.18</td>
<td>27.9</td>
</tr>
</tbody>
</table>

Table 3. Statistics and resulting model precision for each student cluster, with totals and averages in bottom row.

4 when examining discrepancies between in-state and out-of-state admissions on public universities.

Figure 4. Comparison of in-state vs out-of-state admission rates for public universities.

Public universities tend to accept more in state students. Problematically, the linear trend demonstrates there is only minor differences between in state and out of state admission rates in our model.

6.3. MLP Analysis

The general trend of the novel clustering method was that higher-performing students resulted in better performance. If this trend applied to the baseline MLP as well, there could be even higher prediction accuracies then those seen in Table 3. In order to investigate, the baseline MLP’s predictions were split based on GPA, test score, and class rank. The model predicted students with a GPA above 3.7, class rank of top 13% or better, and normalized ACT score above 28, more accurately, with a precision 10-12% greater than the remainder of the data set. A more detailed description of the models performance is shown in Figure 5. Student GPA, test score and class rank are normalized to between 0 and 1 and then averaged to produce a metric by which the students in the test set can be sorted. A sliding window of 200 students is then used to evaluate the accuracy of the model at each position in the ranking of this metric. The graph shown is the average of this result across 10 unique MLPs and train/test sets.
6.4. Admission Rates Comparison

The MLP’s performance was also evaluated by comparing the results to the ground truth of university admission rates. Predictions from the baseline MLP were divided by university. Experimental admission rates were calculated with the corresponding accepts and rejects. Actual admission rates were directly from CollegeData.com and reported by the university. One flaw in the data set is there are significant amount of universities with only a few student profiles. Therefore, in this comparison, only universities with greater than 25 student profiles were used. The results are shown in Figure 6.

6.5. Recommendations for Future Research

Two known significant factors in college admissions are essays and extracurriculars. Through techniques involving NLPs, assigning a quality rank of essays to student profiles could dramatically improve the model. Similarly, for extracurricular activities, quantifying the number of extracurriculars and their respective weights would be a significant feature. The amount of data available and amount lost due to cleaning was also an inhibitor to the success of our model. A worthwhile pivot would involve training individual models for each university, and using FOIA requests to accumulate enough data to do so. It is also evident that it is more difficult to predict the admissions of lower performing students. Focusing on optimizing the model performance on this specific subset of applicants would be a necessary improvement.

7. Conclusions

We suggest that there is an upper bound on the accuracy of predicting college admissions due to the importance of features such as essays and extracurriculars. The baseline MLP’s poor accuracy on lower performing students suggests that these factors are more significant in admitting a student with worse grades. As mentioned in recommendations for future research, quantifying extracurricular data and the quality of essays would be an important next step to improve the accuracy of the overall model.

References