An enrolment admission strategy based on data analytics

Michel C. Desmarais

Polytechnique Montreal, Canada michel.desmarais@polymtl.ca
http://www.professeurs.polymtl.ca/michel.desmarais/

Abstract. Every university program that has a limited capacity of enrolment faces the task of selecting the candidates that have the best chance of success. We introduce a selection strategy based on data analytics that only requires a ranking of candidates from different sources to determine a number of candidates to select from each source. The strategy relies on the distribution of student marks and on historical data of each source. It consists in determining a minimal threshold mark which, in turn, is used to determine proportions of students to admit from each source. The strategy ensures a maximum success rate under certain assumptions.

Keywords: Student Enrolment · Learning Analytics · Candidate Selection

1 Introduction

A case for the use of Learning Analytics in educational institutions can be made for the objective of selecting the candidates that have the best chance of success at a given university program. In the words of [8], we can consider the selection process as a standard machine learning prediction task:

> "Admission is to a great extent a prediction task, where admissions committees aim at estimating a candidate's chance of future study success. For these kinds of tasks, Meehl (1954) provided strong evidence for the superiority of the statistical approach over the clinical one. Since then, a plethora of studies has challenged this result but none contradicted Meehl's conclusion (Kahneman, 2011)."

While the candidate selection problem is trivial if the decision is based on a single criterion, such as the result of an admission test score (GPA, for eg., [1]; or GRE), or on any single score by which a candidate can be ranked, such score is not always available. Often, the decision must rely on a set of scores that are not comparable.

The typical situation is that an admission decision is based on the ranking of students within a given cohort and for a given institution. The choice is simple for the students from the same institution, but not for the students from different

institutions. One solution is to ask candidate students to take an admission exam, but this is unpractical for students that apply from abroad or from distant locations. Moreover, the admission test may not be highly reliable [6].

Other solutions, often considered more reliable, are to to revert to interviews and personal statements [2]. But not only are their reliability questioned [4, 5, 7], these approaches also incur issues of time and efforts, which can be critical for large cohorts.

We introduce a means to decide on student admission based on historical data of the host institution itself. Given the information on student marks and their origin, one approach consists in determining the proportion of students from a given origin that are above a given score. The approach relies on computing the expected mean score of a proportion of students above a given score for a given origin. And the key to the approach is that the scores of all students are on the same scale, namely the institution's own grades.

The strategy is first described below, followed by a short demonstration of the impact it has compared to a simpler solution.

2 Historical data cutoff admission treshold (HDCT)

We will refer to the proposed approach as the Historical data cutoff admission treshold (HDCT). To illustrate its basic principle, consider Figure 1. It shows a distribution of student scores on a Z-scale that follows a Normal distribution ($\mathcal{N}(0, 1)$) along with the proportion of students above the score, which corresponds to one minus the *cumulative distribution function* (labeled "cummul. admiss."). The dotted line indicates that the score of 0 corresponds to a proportion of 50% of students are above that score. We can also see from the "cummul. admiss." curve that at score $Z = 0.5$ we have about 80% of students above that score.

This graph is the basis of the HDCT admission process. The general principle is to determine the proportion of students to retain based on a common minimal score, obtained from the institution's historical data. Given that it is reasonable to assume that all student scores are on the same scale, namely the institution's historical scores, they are comparable even though the students may have different origins. And the key is not to rest the decision on a score obtained from the origin institution, but on historical data from the host institution. This approach incurs that the institution keeps track of which origin institution the student comes from and, as we discuss later, of the ratio of admitted students over the number of candidates.

To illustrate the general approach based on the above introduction, figure 2 shows the case where we have students from three different origins, source *a*, *b*, and c. The mean, standard deviation (s.d.), and the relative proportion of

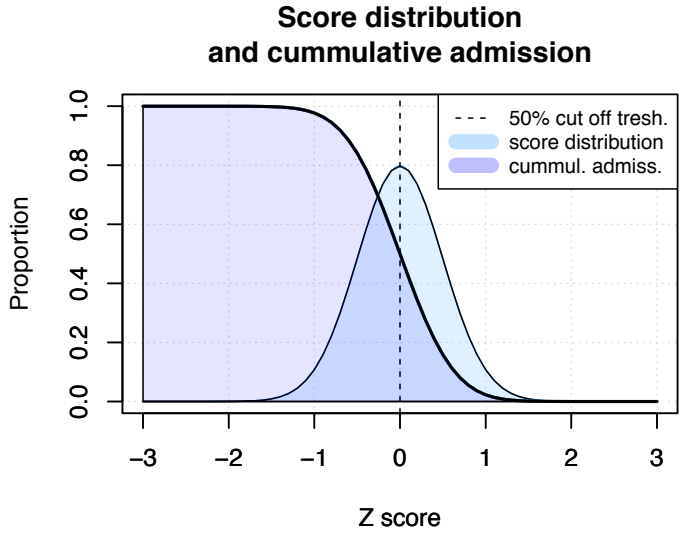

Fig. 1. Relation between the students score distribution on a Z-scale and the cumulative proportion of students admitted.

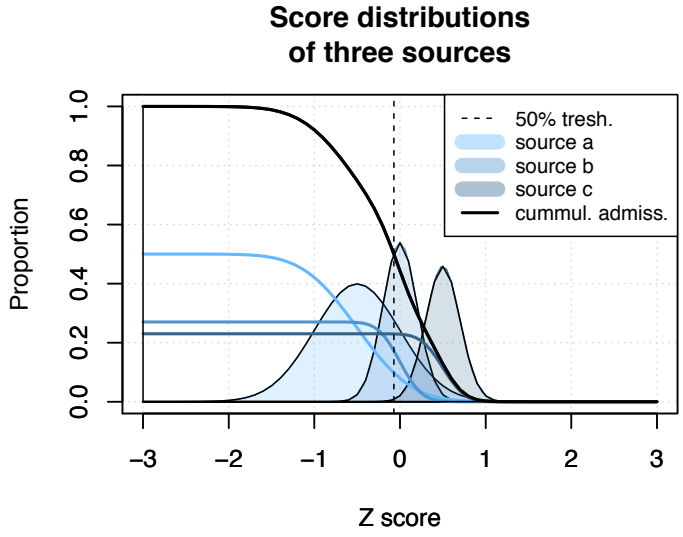

Fig. 2. Score distributions of three sources with different means and sandard deviations. The cumulative admission curve is shown for each source ($1 - cdf$). They correspond to the three colored lines. The global cumulative admission is the black curve and corresponds to the sum of all three cumulative admission curves.

candidates from each source is shown below, along with the proportion admitted at the 50% cut off threshold.

Source	*Mean*	*S.d.*	*Proportion*		*Admitted@50%*	
a	−0.5	0.5	50	%	20	%
b	0	0.2	27	%	64	%
c	0.5	0.2	23	%	99.8	%

We can see from the cumulative distribution curves that source a (mean=−0.5), source-b (mean=0), and source c (mean=0.5) respectively represent 50%, 27%, 23% of all students applicants. Because the variance of the distributions is not equal (0.5, 0.2, 0.2) and they also have uneven proportions, the cut off threshold to admit 50% of students is not at $Z = 0$, but instead around $Z = 0.07$. This threshold is shown as the dotted line in Figure 2: the score where the global cumulative distribution curve reaches 50% of all students, which in turns corresponds to 20% of source a, 64% of source b, and almost all of source c.

The implication of this graph is that if we had, for example, 1000 candidates and we wanted to admit only 500 of them, then only about 200 source a would be admitted, because it has had on average 0.5 standard deviation below the mean in the historical data. Whereas based on a policy of admitting the same ratio for all sources, we would then admit 250 of them for source a. Divergence from a uniform admittance ratio is even more stringent for the other two sources: almost all students from source c would be admitted because they historically scored 0.5 standard deviation above average and have a lower standard deviation, and most of source b would also be accepted.

The Z-score corresponding to the proportion of students we wish to admit from the total applicants is calculated based on an optimization function that can be defined as:

$$\arg\min_{Z} = \sum_{s} \left((prop_{s \in \text{source}} \cdot pnorm(Z, \overline{sco}_s, sd_s)) - prop.admitted \right)^2$$

where:

- $prop_{s \in \text{source}}$ is the proportion of applicants from a given source, s,
- $pnorm$ is the cumulative distribution function (for the Normal distribution) that takes as arguments:
  - Z: the Z-score to optimize (threshold),
  - $\overline{sco}_s$: the mean historical score of the given source, and
  - sd_s: its standard deviation;
- *prop.admitted*: the proportion of students we wish to admit to meet the limited admission capacity.

2.1 Smoothing factor

In some cases, the number of students from a give source may be small, or even nonexistent if it represents a new source. To avoid extreme values of mean and

standard deviations that result from small samples, a smoothing factor should be used. Assuming we have N_s students from source s, a smoothing factor α can be used to bring the mean of the score with the following smoothing formula:

$$\hat{x}_{is} = \frac{\sum_i^{N_s} x_i + \alpha\overline{x}}{N_s + \alpha}$$

where $\hat{x}_{is}$ is the smoothed value that should replace the value of the mean and $\overline{x}$ is the general mean of all students. A reasonable value is to have $\alpha = 5$, although the choice is rather arbitrary. A similar smoothing should be applied to the standard deviation based on historical data.

3 Impact example

To assess the impact of the admission strategy over a simpler one, we run a simulation and compare the difference in the expected scores of students admitted with each strategy.

The simpler strategy is to accept an equal proportion of students from each source.

Let us take the numbers from Figure 2 to run a simulation and assume we admit 1000 students. The expected average score from a given source corresponds to the number of students at a given score ($freq(\text{sco})$), proportionally represented by the source's density of the distribution, times the score. This is repeated for each source and divided by the number of students (N) :

$$\mathrm{E}(\overline{\text{sco}}) = \frac{\sum_{s\in\text{source}} \sum_{\text{sco}\in s} freq(\text{sco}) \times \text{sco}}{N}$$

The numbers that correspond to each strategy for each source are reported in the following table:

	Equal proportion		*HDCT*	
Source	N	$\mathrm{E}(\overline{sco})$	N	$\mathrm{E}(\overline{sco})$
a	250	−0.11	98	0.21
b	135	0.16	173	0.12
c	115	0.66	229	0.50
global	500	0.14	500	0.31

The major difference between the equal proportion and the proposed HDCT approach is that much fewer candidates are accepted from source a for the benefit of greater numbers from sources b and c. The effect is that the expected scores from source a increases while it decreases for sources b and c, but the overall effect is an increase in the expected score of 0.17 $(0.31 - 0.14)$.

4 Conclusion

This paper describes a strategy to select the proportion of candidates to admit in order to maximize the expected success rate of the students to a given program. The strategy is based on historical data from the host institution. The advantage of the approach is that it does not require a standardized score across students from different institutions, which is most of the time unavailable unless the candidates are subject to an admission test. Considering that candidates can come from remote location and that running an admission test can involve considerable time and effort, this is a major advantage.

However, the approach has its limitation, the first of which is to have historical data from the different institutions the candidates come from. Often, the sample can be small and a correction in the form of a smoothing factor is proposed to alleviate this issue.

Another limitation is that, as described in this paper, it assumes the distribution of scores is Gaussian. Now, this limitation is not inherent to the general approach. Non Gaussian, or even arbitrary distributions could be handled, but the computations would need to be adapted to the actual distribution.

Finally, another issue is that the distributions have to reflect the scores of the origin institution, which must be derived from the historical data of the accepted candidates in the host institution. As presented in this paper, we assume the historical data is a faithful representation of that distribution, but if the selection is based on a small proportion of applicants, this assumption would be false. Here again, this is not a limitation of the approach itself, and computational adjustments would have to take this factor into account. The adjustment will rely on information about the ratio of admitted students per institution.

To close the loop on the question of how Learning Analytics can bring value to education, we use the admission problem that every institution faced with the need to select candidates from disparate source is confronted with. The candidate selection approach uses a strategy that relies on statistics and optimization techniques. It is an objective, effective, and efficient means to achieve the goal of selection the candidates that have the best chances of success.

References

1. Didier, T., Kreiter, C.D., Buri, R., Solow, C.: Investigating the utility of a GPA institutional adjustment index. Advances in health sciences education **11**(2), 145–153 (2006)
2. Eva, K.W., Reiter, H.I., Rosenfeld, J., Trinh, K., Wood, T.J., Norman, G.R.: Association between a medical school admission process using the multiple mini-interview and national licensing examination scores. Jama **308**(21), 2233–2240 (2012)
3. Kahneman, D.: Thinking, fast and slow. Macmillan (2011)
4. Meehl, P.E.: Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. (1954)
5. Murphy, S.C., Klieger, D.M., Borneman, M.J., Kuncel, N.R.: The predictive power of personal statements in admissions: A meta-analysis and cautionary tale. College and University **84**(4), 83 (2009)
6. Salvatori, P.: Reliability and validity of admissions tools used to select students for the health professions. Advances in Health Sciences Education **6**(2), 159–175 (2001)
7. Siu, E., Reiter, H.I.: Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. Advances in Health Sciences Education **14**(5), 759 (2009)
8. Zimmermann, J., von Davier, A., Heinimann, H.R.: Adaptive admissions process for effective and fair graduate admission. International Journal of Educational Management **31**(4), 540–558 (2017). https://doi.org/10.1108/IJEM-06-2015-0080, `https://doi.org/10.1108/IJEM-06-2015-0080`