TheDataOpen

The Data Open

Data Open Championship - Team 16

Contents

1 Non-Technical Executive Summary

1.1 Background

In this competition, we are provided with a dataset containing the characteristics of postsecondary institutions across the United States. Our objective is to find the quantifiable strengths and weaknesses of this system.

1.2 Exploratory Data Analysis

Aggregated ethnicity distribution Let us look at the ethnicity distribution for the fall enrollment and compare it to the associated graduation rate. The ethnicity proportions are quite stable between enrollment and graduation, with a slight increase in the graduation proportion for the white ethnicity, while the rest seems to be decreasing slightly. One first observation can postulate that minorities tend to drop out more than white majority.

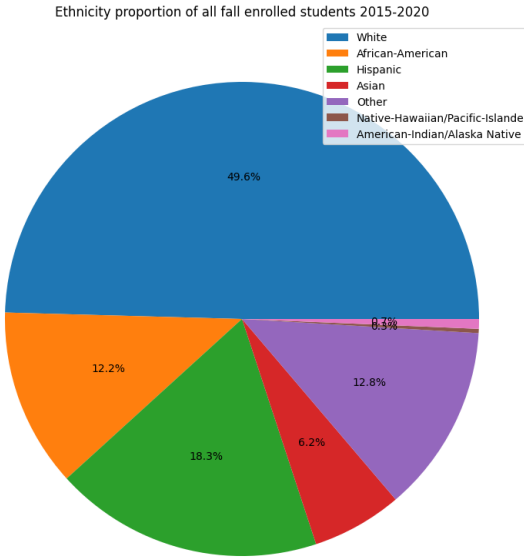

Figure 1: Ethnicity distribution overall for fall enrollment (2015-2020)

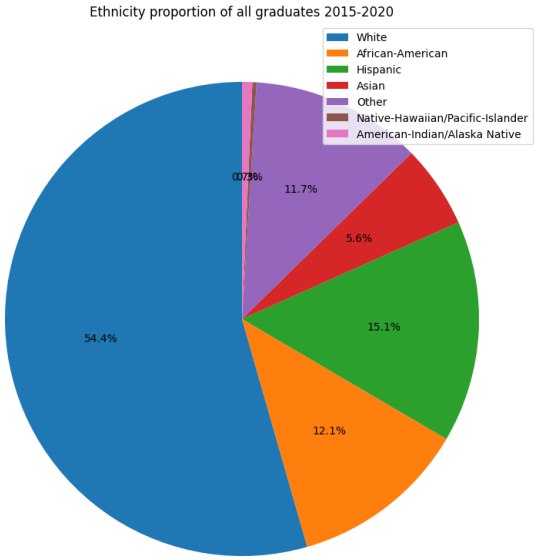

Figure 2: Ethnicity distribution overall for graduates (2015-2020)

Granular institution level ethnicity distribution Another thing to look at is the distribution of ethnicity among the top US institution in terms of number of students enrolled. The first university in terms of total number of students in this visualization is actually an outlier compared to its peers as it has a majority of online learning students, hence the disparity in collecting ethnicity information.

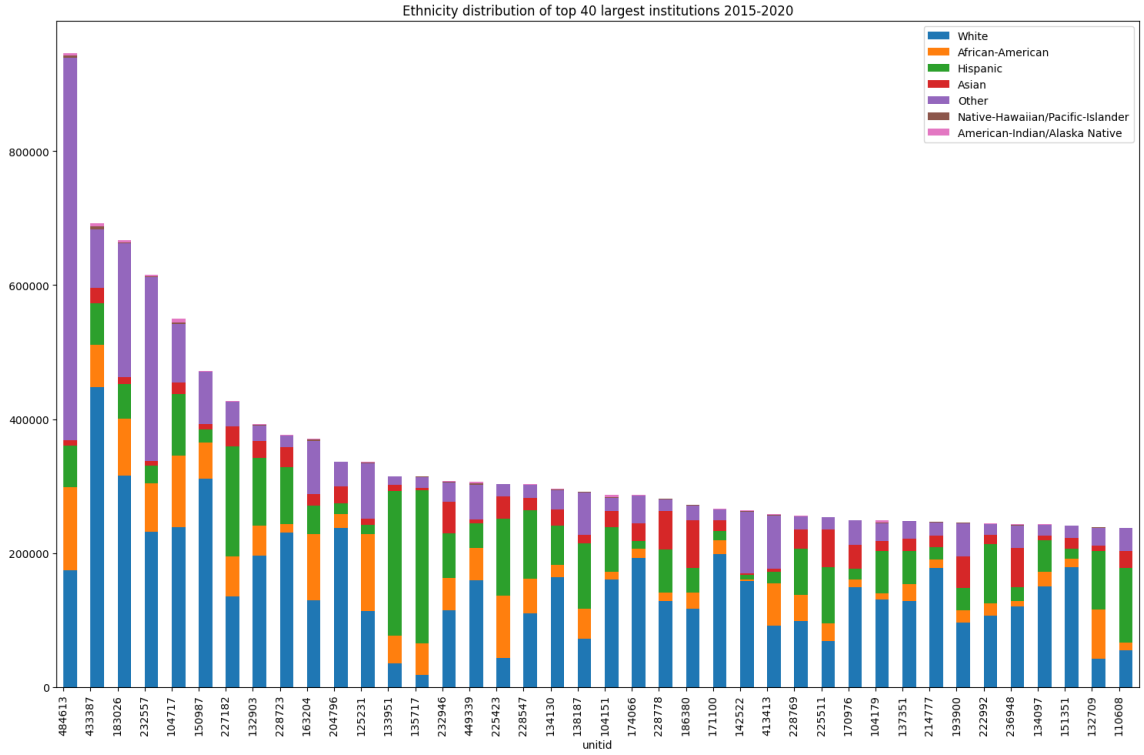

Figure 3: Distribution of ethnicities per institution

Proportion of Tuition Fees in Total Revenues Institutions require adequate funding to provide quality education and research. Now, let us inspect the distribution of how much tuition fee revenues account for total revenues. For example, a value of 0.5 suggests that tuition fee counts for half of the institution's total revenue. As shown in the histogram below, a rough normal distribution with an approximate mean value of 0.6 is seen. This suggests that on average, tuition fees account for a significant amount of revenue which the institution must use for teaching, research and more.

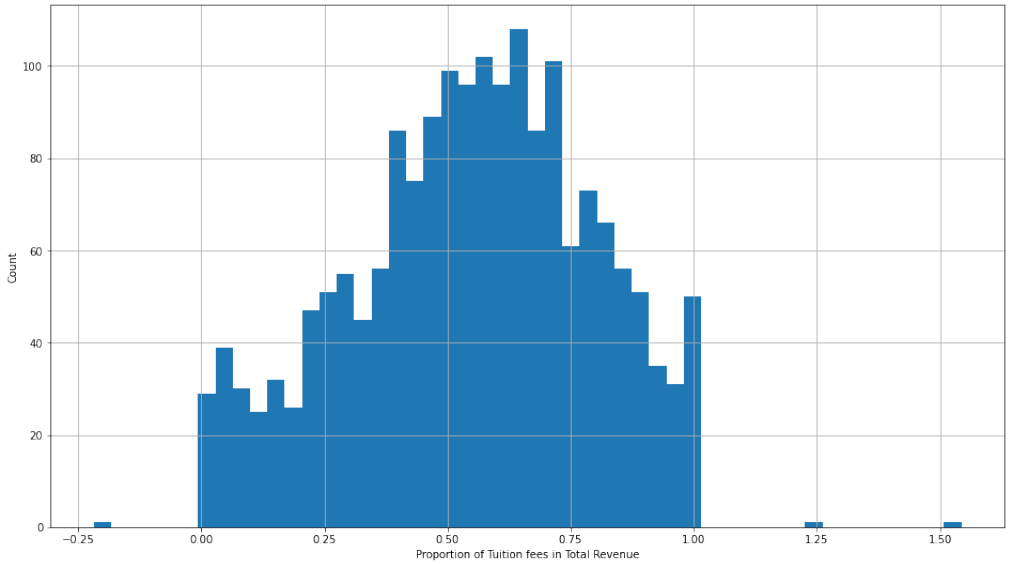

Figure 4: Distribution of Tuition fees per institution

1.3 Problem Formulation

Thanks to EDA, we managed to come up with the following question that we will try to answer in this project: Can we create a high-quality ranking of postsecondary institution using a mathematical model?

Motivation: Today's education rankings are based on merit and prestige, however we believe in a more flexible and accessible approach in ranking postsecondary institutions.

2 Modelling

Our goal is to compute define and to compute the effectiveness of the postsecondary education system in the United States. Then, our goal was to see if this system can be improved with respect to our metric. To do so, we first selected parameters that could explain the values of our metric (e.g. the ethnicity proportions in each school). Then, we conceived a model to compute the values of our metric for new values of the parameter, i.e. we parametrized the effectiveness metric according to one or several parameters.
Finally, we run an optimisation algorithm to optimize the effectiveness of the postsecondary education system, and we computed several stastics to confirm or reject the reliability and the fairness or our new market design.

Our methodolody is quite robust and can be adapted to other effectivness metrics and other parameters to calibrate. Due to the time constraint, we have made several

assumptions and simplifications to define the system effectiveness metric and to compute it. We also ran the optimisation algorithm for a few number of iterations due to the long computational time of the metric.

2.1 Postsecondary Education System Effectiveness

Let $n \geq 2$ be the number of postsecondary instututions in our study. We denote them by $\mathcal{Q}$ the set of insitutions. Our goal is to define a parametric *quality measure* of any institution $q \in \mathcal{Q}$ that we denote $\pi(q)$. Then, we define the effectiveness of the system $\mathcal{Q}$ by:

$$\pi(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \pi(q).$$

Assumptions - For this initial model, we assume that a high quality institution has:

- ⋆ A diverse student body
- ⋆ A high graduation rate
- ⋆ Low tuition fees
- ⋆ A low selectivity rate (i.e. the university is open to all and is able to educate any student, regardless of their initial academic background)

Hence, in our model, the quality of a university is defined as a function of the student body diversity, the graduation rate, the tuition fees and the selectivity rate of the university. Notice that many other parameters could have been added to this definition, like the quality of the teaching staff or the number of libraries per university etc. Nonetheless, as a first step, we have defined the above metric that encompasses several characteristics of a 'good education': the academic quality of the entering students, the cost (for the students), the quality of the education (reflected in the graduation rate) and the diversity of profiles that make up the student cohort.
From these variables, the idea is to define the university quality by:

$$\frac{\text{student body diversity indicator} \cdot \text{graduation rate}}{\text{tuition fees intensity} \cdot \text{selectivity rate}}$$

Now, we need to compute (or at least approximate) the value of each of these parameters to define the quality of a university.

2.1.1 Student body diversity

From the file `EFA_2015-2020_data.csv` in the folder `Fall Enrollment`, we can compute the average proportion of each ethnicity in each school from 2015 to 2020. More precisely, we have $e = 7$ classes (American Indian or Alaska native, Asian, Black or African American, Hispanic, native Hawaiian, White or other [two or more races, not defined]). For each institution $q \in \mathcal{Q}$, we compute $p_1^{(q)}, p_2^{(q)}, ..., p_e^{(q)}$ the average proportion of student belonging to each ethnic class (averaged from 2015 to 2020). These simple statistics are used to describe the racial makeup of each institution.

To compute the diversity of an institution, an intuitive idea would be to check if the proportions of each ethnicity are close to each other (i.e. $\approx \frac{1}{e}$). In that case, the diversity indicator would be maximal. On the contrary, if one ethnicity is over-represented, then the diversity indicator would be close to zero. An intuitive way to build a such metric would be to consider the standardized Shannon entropy of a sample $p_1^{(q)}, p_2^{(q)}, ..., p_e^{(q)}$:

$$\frac{1}{\log(e)} \sum_{k=1}^{e} p_k^{(q)} \log\left(\frac{1}{p_k^{(q)}}\right) \in [0, 1].$$

However, some ethnicities are naturally under or over-represented in the total population. So instead, we decided to compare $p_1^{(q)}, p_2^{(q)}, ..., p_e^{(q)}$ to the average proportion of each ethnicities in the full cohort $\overline{p_1}, \overline{p_2}, ..., \overline{p_e}$, and we computed the Jensen-Shannon distance between the two samples:

$$\Delta(q) := \sqrt{\frac{\mathcal{D}(\left(\overline{p_1}, ..., \overline{p_e}\right)||(p_1^{(q)}, ..., p_e^{(q)})\right) + \mathcal{D}\left((p_1^{(q)}, ..., p_e^{(q)})||(\overline{p_1}, ..., \overline{p_e})\right)}{2}} \in [0, 1],$$

with $\mathbf{D}(p||q)$ the Kullback-Leibler divergence defined by:

$$\mathbf{D}(p||q) := \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

which is used to measure the expected excess 'surprise' of getting specific ethnic proportions compared to what is observed in the full cohort.

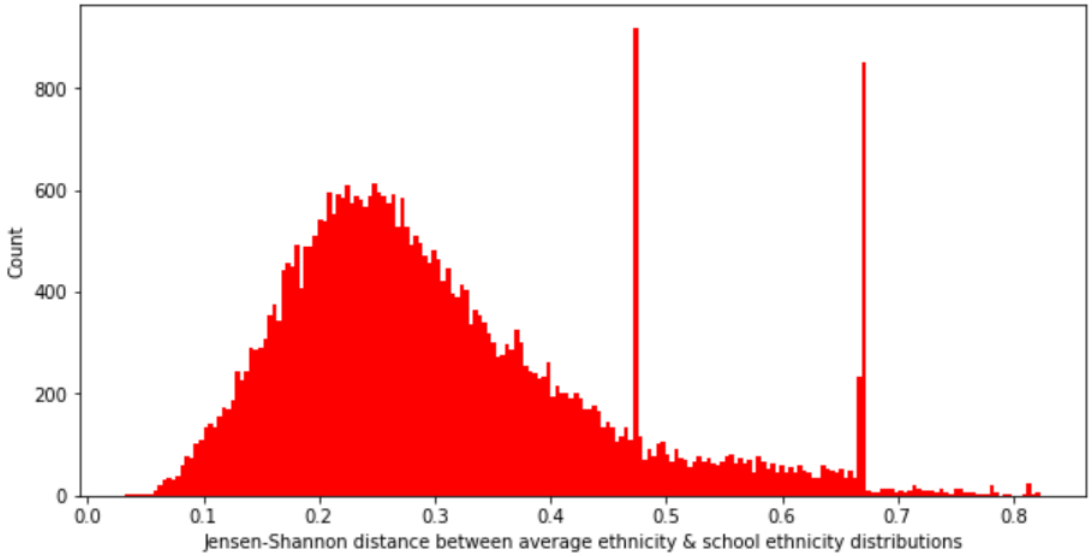

Figure 5: Diversity scores distribution (2015-2020)

We observe that most of the diversity scores are centered around 0.25, which is not too far from the overall distribution of the cohort. However, we observe two peaks in the distribution around 0.48 and 0.68, meaning that a large number of schools are not truly diverse.

2.1.2 Graduation rate

We estimated the graduation rate of each university by using the file `OM_2015-2021_data.csv` in the `Outcome Measures` folder. More precisely, we approximated the graduation rate by the column `OMAWDP4` that represents the proportion of students who graduated with a degree (associate or bachelors) or a certificate after 2-4 years of education.

We used this surrogate because most of the students complete their degree in 2-4 years, and computing the graduation rate with the other students would have been much more complicated (because the enrollment date is a function of the duration of the program). To simplify the computation, we restricted our metric to the undergraduate students.

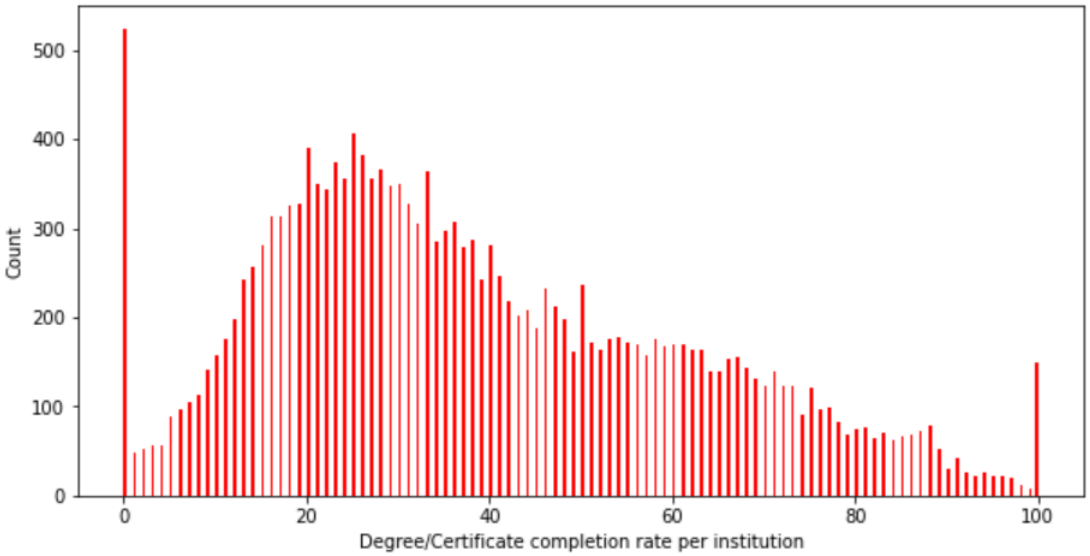

Figure 6: Graduation rates distribution

2.1.3 Tuition fees intensity

The idea of this metric is to measure the profit made by the university thanks to the tuition fees. The intuition is that education should be as cheap as possible for students, and a university should be able to provide a good education to its students without charging too many tuition fees. The indicator is close to 1 if the institution charges a lot of tuition fees compared to the others, and 0 otherwise.

More precisely, for each institution, we used the file `F_F2_1415-1920_data.csv` in the `Institutional Finances` folder to compute the total profit of the university made on tuition fees (column `f2d01`). It gives amounts $(a_q)_{q\in\mathcal{Q}}$. To get the tuition fees intensity, we computed the cumulative distribution function of these numbers:

$$\forall q \in \mathcal{Q}, \qquad \mathbf{F}_{\text{fees}}(a_q) = \frac{1}{|\mathcal{Q}|}\sum_{q'\in\mathcal{Q}}^{n} \mathbf{1}_{(a_q \leq a_{q'})}.$$

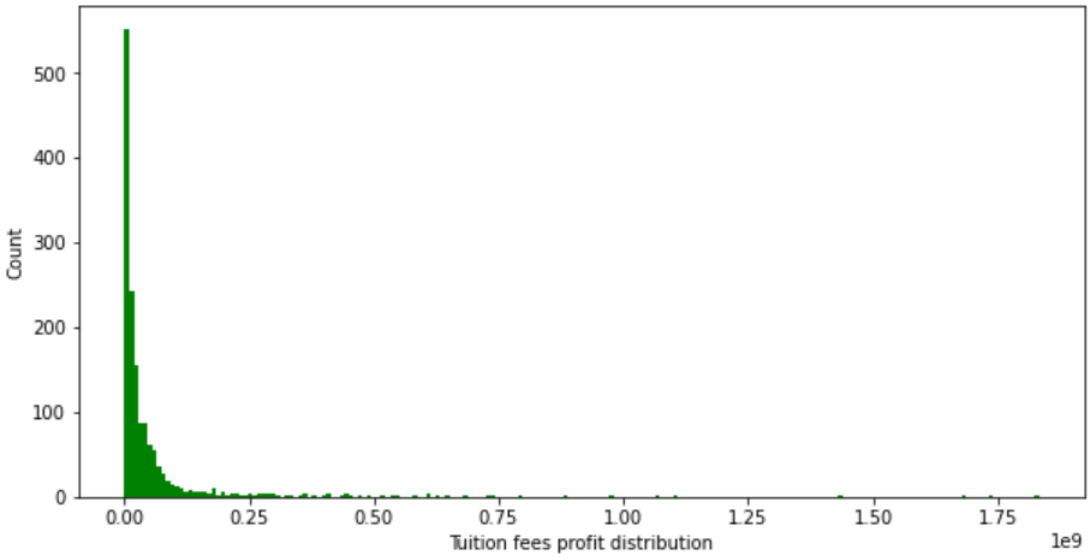

Figure 7: Tuition fees profit distribution

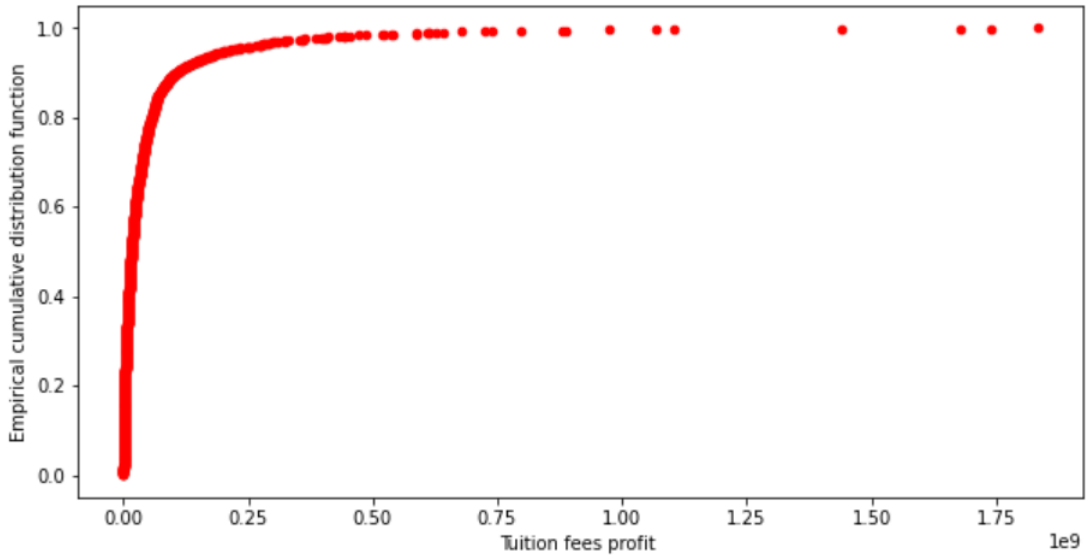

Figure 8: Empirical cumulative distribution function of the tuition fees profit

2.1.4 Selection rate

In a similar way to part 1, we approximate the selection rate of an institution by its admission rate. To compute it, we used the file `ADM_2015-2021_data.csv` in the `Admission and Test Scores` folder. More precisely, we used the columns `aaplcn` and `admssn` to compute the ratio between the total number of admissions over the total number of applicants in each institution. For the institutions without data, we approximated the ratio by the median of our sample.

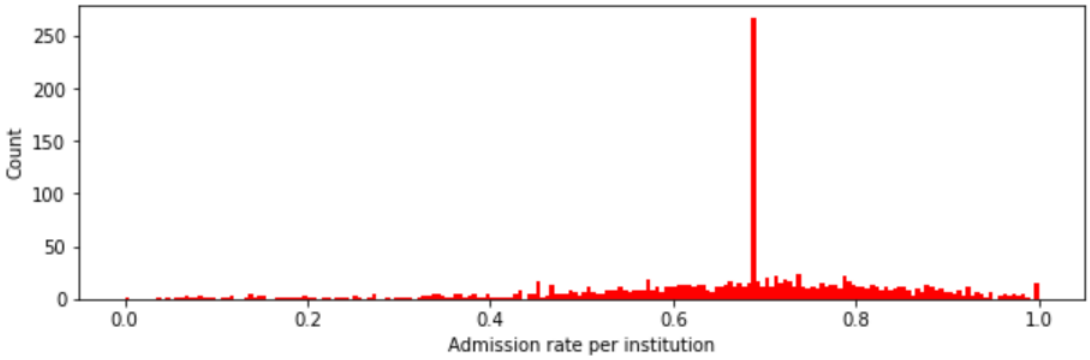

Figure 9: Admission rates distribution

2.2 Education System Effectiveness: Formal definition

Based on our previous metrics, we define the *quality* of an institution $q \in \mathcal{Q}$ by:

$$\pi(Q) := \frac{\left(1 - \Delta\left(p_1^{(q)}, ..., p_e^{(q)}\right)\right) g(q)}{\mathbf{F}_{\text{fee}}(q) s(q)}$$

with $g(q)$ and $s(q)$ the graduation and selection rates defined above.

The *effectiveness of the education system* is defined by:

$$\pi(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \pi(q).$$

2.3 Education System Effectiveness: Analysis

We computed the quality scores for each institution and plotted the distribution of these scores:

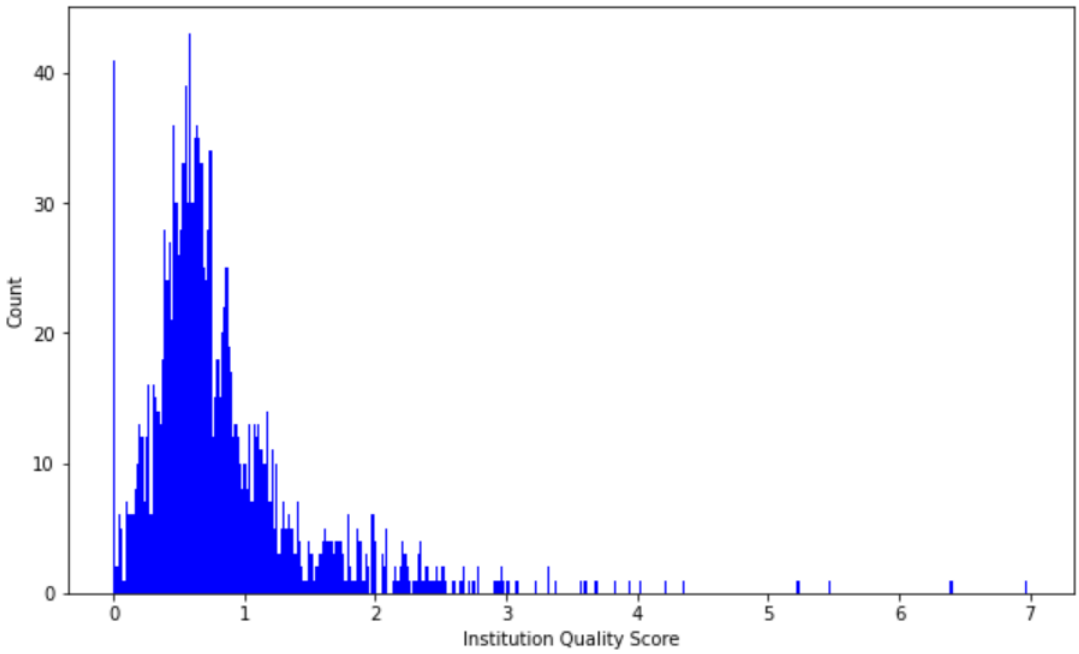

Figure 10: Institution quality scores distribution

Our model suggests that the institution quality score follows a kind of exponential distribution: most of them have a low quality score, due to high tuition fees, very low selection rates, low graduation rates or low diversity scores. We observe that only a few number of institutions are able to get top quality scores.

Model Results Below is the table of the top ranked schools using this model. It is interesting to note that the schools listed here are not well-known, as the famous target schools or Ivy League schools are very selective, which lowers their quality score.

UnitID	Name
# 183600	Assumption College for Sisters
# 211893	Curtis Institute of Music
# 495280	Indian Bible College Flagstaff
# 437705	Monteclaro Escuela de Hoteleria y Artes Culinarias
# 181543	Summit Christian College
# 446613	W L Bonner College
# 197221	Webb Institute

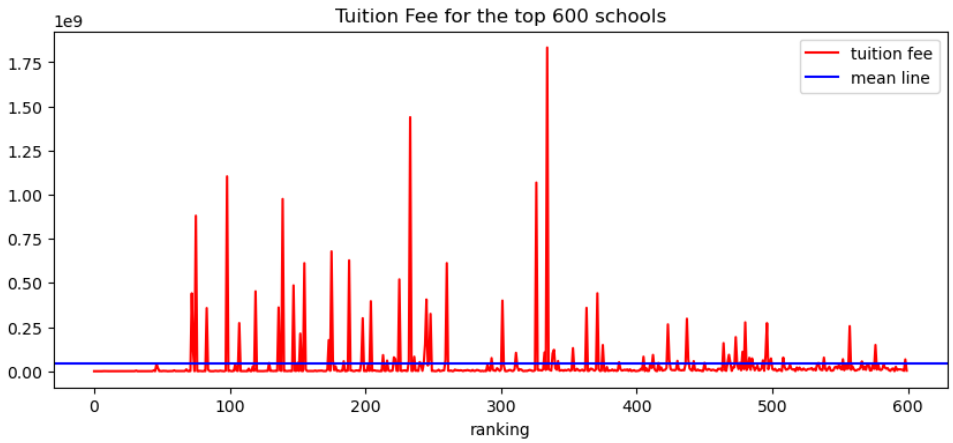

Figure 11: Tuition fee distribution for the top 600 schools

Here, we have the tuition fee distribution for the top 600 ranked schools, ranked by our model. Clearly we can see that the top 100 are much less than the overall mean, with the exception of a few outliers. The tuition fees are quite noisy subsequently but are still higher than the mean. Our model has indeed tried to quantify low tuition fee as a better feature, which seems to have worked for the top rankings.

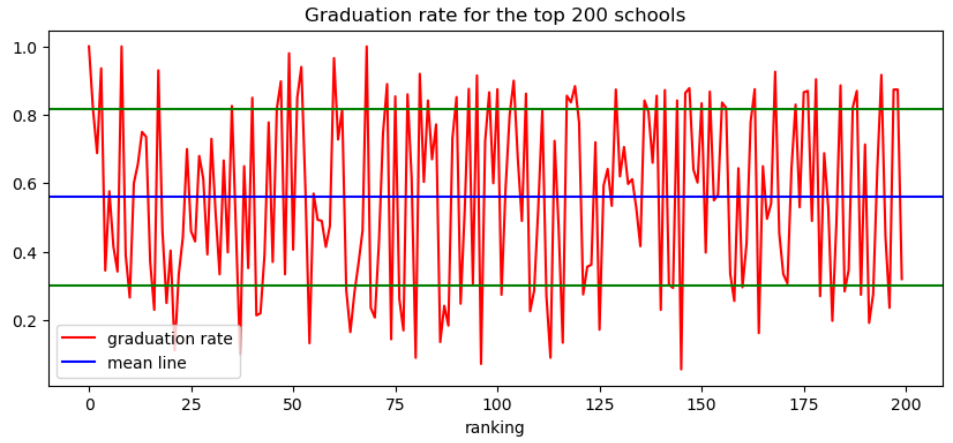

Figure 12: Graduation rate for the top 600 schools

The graduation rate shows a slight downward trend from the first 25 ranked schools but is very random afterwards. This shows graduation rate does not impact the ranking of the school much, except the top ones.

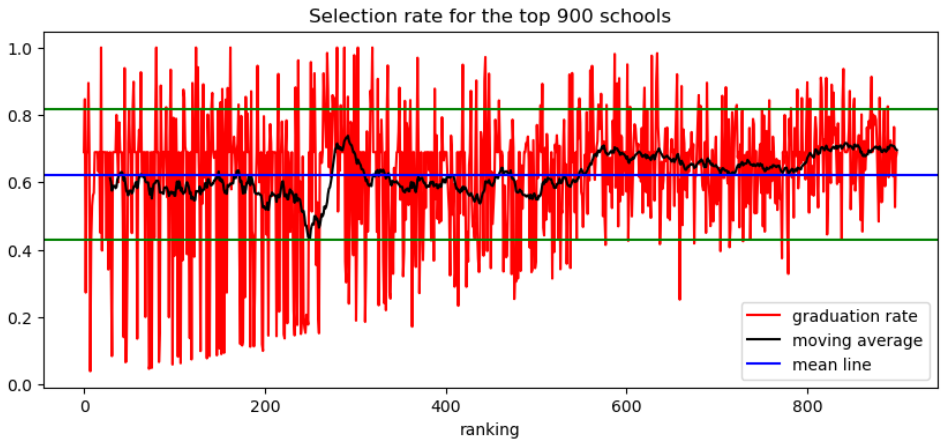

Figure 13: Selection rate for the top 900 schools

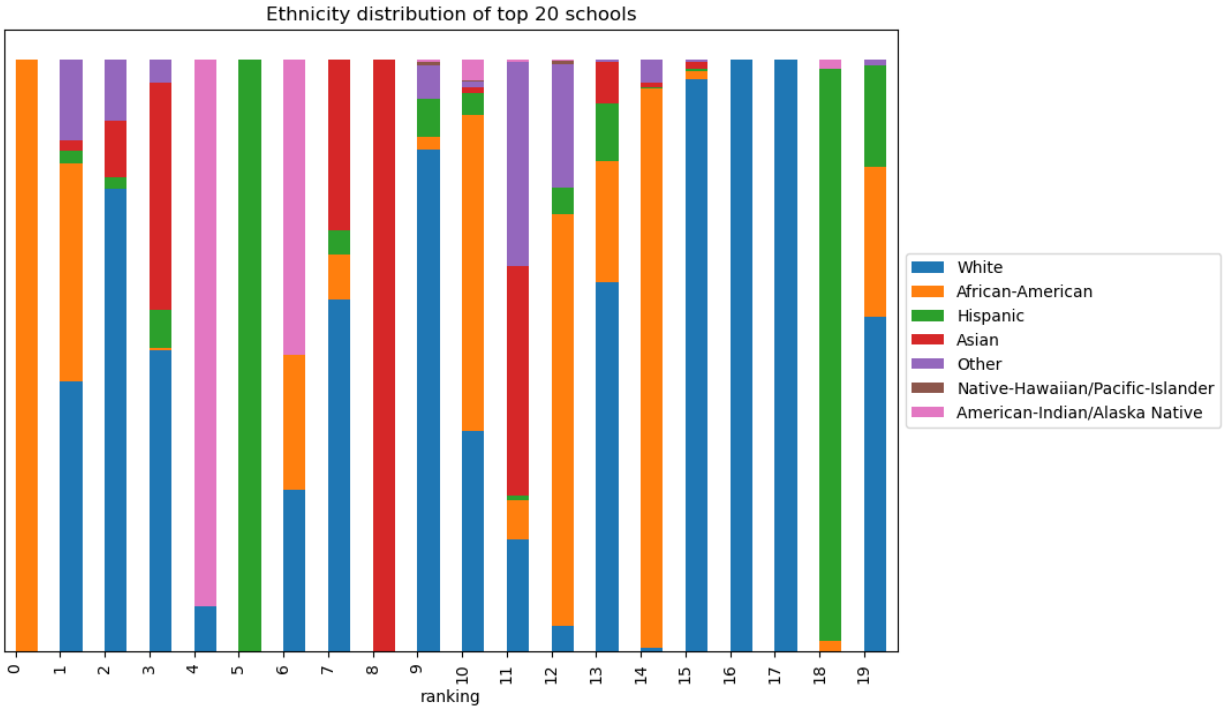

Figure 14: Ethnicity distribution of the top 10 schools

Finally, the ethnicity distribution for the top 10 most effective schools look completely random and not at all optimized for diversity, at least on a granular level (per institution, as some institutions are uniquely ethnic (e.g. the first one with African-American, or the 15th through 17th with White). Despite the lack of diversity on the top scorers, its other high-scoring features pull up its ranking. Still, diversity has room for more improvement and is something to look further into– it is not an easy problem to solve.

3 Education System Effectiveness: Optimisation

Now, we would like to see if some parameters of the market design could be modified to improve the effectiveness of the education system $\pi(\mathcal{Q})$. For example, one could be interested in setting quotas or reserves in each school for disadvantaged or underrepresented minorities to improve diversity in each school and prevent segregation. However, our model is not dynamic and we are unable to compute the education system effectiveness for unknown values of our parameters.

To set up our experiment, let's say that we want to calibrate the proportion of each ethnicity in each school (i.e. the $p_1^{(q)}, ..., p_e^{(q)}$ for each $q \in \mathcal{Q}$ to optimize the system effectiveness. To do so, we can assume that for each school $q \in \mathcal{Q}$, the graduation rate and the selection rate are functions of $p_1^{(q)}, ..., p_e^{(q)}$, i.e.

$$g(q) := f_1(p_1^{(q)}, ..., p_e^{(q)})$$

and

$$s(q) := f_2(p_1^{(q)}, ..., p_e^{(q)})$$

to find these functions, we train machine learning models (here the boosting algorithm XGBoost) on the whole dataset. However, by training these models, we observe that their performances are very low ($R^2 \sim 25\%$ for f_1 and $< 5\%$ for f_2), which means that the selection rate and the graduation rate might not be influenced/caused by the ethnicity proportions in each school. Hence, we keep these values constant (and same for the tuition fees intensity: we assume that students will still pay their tuition fees, regardless of their ethnicities.)

Hence, we want to find:

$$\left(p_1^{(q)*}, ..., p_e^{(q)*}\right)_{q\in\mathcal{Q}} = \underset{\left(p_1^{(q)}, ..., p_e^{(q)}\right)_{q\in\mathcal{Q}}}{\text{Argmin}} \; - \frac{1}{|\mathcal{Q}|} \sum_{q\in\mathcal{Q}} \frac{\left(1 - \Delta\left(p_1^{(q)}, ..., p_e^{(q)}\right)\right) g(q)}{\mathbf{F}_{\text{fee}}(q) s(q)}$$

In this case, since all the constraints are independent for each school, can minimize each term of the sum individually. Each term is maximized when $\Delta\left(p_1^{(q)}, ..., p_e^{(q)}\right) = 1$, i.e. $(p_1^{(q)}, ..., p_e^{(q)}) = (\overline{p_1}, \overline{p_2}, ..., \overline{p_e})$, i.e. when ethnicity proportions in each school represent the average ethnic makeup of the undergraduate students.

4 Education System Effectiveness: Optimisation results & Extended tests

We have implemented the optimisation routine in Python using `Scipy`. However, the code takes a lot of time to run (computing the Jacobian matrix takes a lot of time because there are $(|\mathcal{Q}| \cdot e)^2$ parameters to compute at each iteration) so we are currently not able to plot the output of the optimisation process. Then, we decided to move on to `Gurobi` to try and solve this more efficiently. Again, the computation of this problem were out of reach. However, here are the statistics that we planned to compute to measure the efficiency of our solution:

- ⋆ Check that the education system effectiveness is improved compared to the current one
- ⋆ Check if the which schools are the highest changes in ethnicity proportions (high quality or low quality schools?). If high or low quality schools are impacted by these changes, it would prove that there is still some racial segregation in some schools and that it affects the overall efficiency of the postsecondary education system in the US.
- ⋆ Check which ethnicities are have the highest average (%) change overall all the schools.
- ⋆ Check the location of the schools with the highest changes.

5 Conclusion

In this project, we deep-dived into a few of the data sets in order to find quantifiable strengths and weaknesses of this system including: Ethnic Diversity, Graduation Rates, Selection Rate, and Tuition Fees.

We proposed a mathematical model to find an effectiveness level(metric) for undergraduate institutions in the United States, that rewards diversity and graduation rates and penalises higher tuition fees and selectivity.

XGBoost was used to fit ethnicity data with graduation rates and selection rates, however their respective models scored poorly. This showed that selection rates and graduation rates are unlikely to be influenced by the proportions of ethnicity in each institution.

We kept these values constant along with tuition fees and formulated a minimised constraints problem to maximise the effectiveness score for every institution in the education system. This optimisation process would take too long to compute given the short time-frame of the competition. Despite this, we still can apply a mathematical model of effectiveness to compare and rank institutions.

Possible Avenues With more time, we could deploy a classification task on whether students have gotten financial aid or not, and try to ascertain any potential bias between majority and minority class (for e.g. White v. Other, or Men v. Women). There exists a variety of methods to do this bias detection, such as the α-gap or optimal classification trees which split on the protected class. To alleviate any potential bias, we can use False Positive Rate (FPR) control using mixed integer programming, allowing us to identify fairer financial aid distribution. Such methods could be reused and applied to other variables, such as admission in itself.

Hence, we could use this analysis and produce a personalised ranking for a high school student applying to college. This ranking is tailored to the student's preferences such as prestige (selectivity), cost of fees, library funding, graduation rate, diversity and more. These variables can be rewarded/penalised in the quality model $\pi(Q)$, reusing the metrics as calculated in 2.1. The final product would ensure convenience and satisfaction in the next step of the student's education.