Modeling The Loss of Diversity at Harvard College in a Post-Affirmative Action United States

Trey Whitehead '25
Harvard College
treywhitehead
@college.harvard.edu

Liam Bieber '25
Harvard College
liambieber
@college.harvard.edu

Sophie O'Melia '25
Harvard College
sophieomelia
@college.harvard.edu

***Abstract*—Given the present Supreme Court case revolving around the legality of Affirmative Action in the United States, we [the authors] decided to develop a post-Affirmative Action, automated college admissions system for our school, Harvard College. We designed and calibrated a stochastic model that can predict the number of admitted students by race. However, race has no impact on the evolution of the dynamics of the system itself. Rather, it is assigned as a byproduct of household income, and the model takes several other non-racial factors as inputs. These factors include, but are not limited to academic performance, geographic origin, and extracurricular activities. Each specific characteristic has been weighted individually, so that when a candidate receives their rating in each area, a weighted average can be calculated to determine the "strength" rating of their holistic application. This rating is compared to a cut-off threshold, which determines acceptances and rejections in a binary manner. We have found that our model, when calibrated with correctly, admits students at a comparable rate to reality. To showcase these results, we have compare our results to Harvard's recently admitted class.**

I. Introduction

"You can't seriously suggest that demographics aren't a factor to be looked at in combination with how isolated–or not isolated–your student body is actually reporting itself to feel?"[1]

Justice Sonia Sotomayor posed this rhetorical question in an oral argument over Fisher vs. University of Texas at Austin in 2016. The landmark Supreme Court case targeted the university's use of race in its admissions process. Abigail Fisher, a white applicant, claimed that she was denied admission because of her race, while the university argued that race was one of many factors used to create a diverse student body. The Supreme Court ultimately ruled in favor of the University of Texas, upholding the constitutionality of affirmative action in higher education, but also stating that race-conscious admissions policies must be narrowly tailored and reviewed for their continued necessity.

However, a new Supreme Court and continued turbulence surrounding the admissions debacle have brought the future of affirmative action into question. Will it be around to stay? If repealed, what might the future of the college admissions process look like? The debate has even entered our own backyard, as Harvard University continues to be engaged with its own Supreme Court case.

In 2014, the Students for Fair Admissions (SFFA), a non-profit organization seeking to change *"unfair"* racial classifications and preferences within the college admissions process, sued Harvard, alleging that the holistic admissions process discriminates against Asian-Americans. In the spring of 2015, another group filed a similar complaint against Harvard with the Department of Education. Although the complaint was dismissed in the following summer, the investigation was reopened under the Trump administration.[2]

As evidenced by the University of Texas case, Harvard is not the only higher education institution under fire for their admissions processes. The SFFA has pointed to the rankings Harvard places on their applicants' likability, courage, and kindness, which they have deemed to be highly-subjective. Furthermore, the SFFA takes issue with the quotas that college sets on different types of students. Whether directly or indirectly, the group claims that each of these practices has discriminated against Asian American students, and they have challenged Harvard to alter or disband its ways.

The outcome of Students for Fair Admissions v. Harvard will determine the future of higher education in the United States. If considering race as a factor of a student's application is prohibited, college admissions will change forever.

It is the goal of this paper to develop a post-Affirmative Action college admissions system in which race cannot be taken into account. Through our research, we intend to quantify and calculate the potential loss in racial diversity at Harvard University.

A. Questions

We intend to leverage our model to gain insight into the following phenomena:

1) *What value of c (the cut-off value) generates a class size most comparable to that of the incoming class of 2027 at Harvard?*

[1]Mears, Bill. "Selected Quotes from Supreme Court Affirmative Action Arguments." CNN, 9 June 2013, www.cnn.com/2013/06/07/politics/court-affirmative-quotes/index.html.

[2]Hlr."Students for Fair Admissions, Inc. v. President amp; Fellows of Harvard College." Harvard Law Review, 24 Mar. 2023, harvardlawreview.org/print/vol-134/students-for-fair-admissions-inc-v-president-and-fellows-of-harvard-college/.

2) *What are the respective yield and acceptance rates for this value of c?*
3) *Which racial group(s) will be hurt most by the repeal of Affirmative Action?*
4) *Which racial group(s) will benefit?*
5) *What does the frequency distribution of ϕ_k-s look like in an admitted and committed student class?*

. . .and most importantly:

To what degree is diversity lost or gained in a post-Affirmative Action system?

B. Disclaimer

It is important to note that the authors of this paper do not endorse either side of the Students for Fair Admissions v. Harvard case. As an ambivalent third party, we strictly hope to provide a framework through which college admission might be analyzed, should Affirmative Action be repealed.

II. The Model

In order to represent what a future Harvard college admissions scheme might look like, we decided to work with a stochastically-driven model. We devised a system from scratch; starting from the demographic composition of a given applicant class and the framework of a college application. These elements included GPA, standardized testing scores, geography, extracurriculars, the personal essay, the interview and legacy status.[3]

The model flowed as follows: firstly, we determined the applicant's income quintile. While constructing each applicant and their attributes, the applicant's race was probabilistically determined as a function of their income using both US Consensus Data and race breakdown of the Harvard Applicant pool. In other words, each median household income quintile can be subdivided by race. Then, we scaled the percentages of each race at each quintile by the percentage of that race which filled the Harvard Applicant pool. This filled out the ethnicity of our applicant pool. Before we go any further, we must stress that although race is probabilistically determined through income for the model, it does not factor into an applicants rating (as it would not in a post-affirmative action world). Instead, an applicant's race is simply stored to compare our final model results to those of Harvard's actual admitted class.

Once both income quintile and race were determined for the applicant, we used primary research to quantify the positive correlation between household income and standardized testing scores.[4] From there, we noted the strong relationship between standardized test scores and academic performance (GPA). We tied the applicant's GPA to their test scores, $+/-$ a variant factor. Then, we accounted for a random multiplier on certain geographic locations (i.e. New York) that send more students to Harvard than say, Kansas. We then stochastically assigned a weighted score for a student's extracurricular activities, their personal essays, and their interview performance. Finally, as Harvard has a reputation for bias towards legacy students, we factored this into our equation.

With all our attributes set, we then developed the notion of a strength vector to quantify the vigor of each applicant's application. Each element within the strength vector was built up as described vaguely above – the exact process will be explicated in the following section. After each applicant's strength vector had been combined, we instituted a cutoff value in order to determine the percentage of the applicant class would achieve admission into the university. This cutoff value was determined empirically in order to generate a class of approximately 1960, a figure in the neighborhood of a standard Harvard class of size. From there, we initialized a series of bins to represent admitted, wait-listed, and denied applicants. Guessing the yield rate to be 90%, we were able to determine the makeup of the incoming class and plot the results. We will now provide a more in-depth examination of the model, alongside the empirical data that helped us arrive at our final figures.

A. Derivation (I)

Applicant Pool Size. Prior to discussing the results of our model, we will rigorously define and defend its foundation. Basing our applicant pool size on the Harvard Class of 2026, we found that $61,221$ had applied. Going forward, Ω will denote the size of a given applicant class.

Rating Vector. Let us first denote two fundamental vectors: the rating vector and the weight vector. The rating vector, $\vec{rating} \in R^{1\times 5}$ encapsulates the model's primary rating system. Its 5 entries are evenly spaced out between 0 and 1 such that $\vec{rating} = [0.0, 0.25, 0.30, 0.75, 1.0]$.

Weight Vector. The weight vector is a product of the application process itself. Although we were unable to find exact data on the degree to which Harvard weights each part of their prospective students' applications, we decided to define our weight vector as follows: $\vec{weight} \in R^{1\times 7}$, where:

$\vec{weight}$ *= [Test Score Weight, GPA Weight, Geographic Weight, Extracurricular Weight, Personal Essay Weight, Interview Weight, Legacy Weight].*

After consulting our former college admissions counselors from our respective high schools, we decided on the following notably subjective values for Harvard:

Please note that all elements within the $\vec{weight}$ must sum to 1. This relationship is imperative to determining an individual candidate's strength rating. Although the exact

[3]It is worthwhile to note that the model does not account for increased admittance percentage for athletes.

[4]*SAT Suite of Assessments Annual Report.* College Board, 2022, https://reports.collegeboard.org/sat-suite-program-results

Weight Type	% Value
Grade Point Average	20%
Standardized Test Score	10%
Geographic Region	15%
Extracurricular Involvement	15%
Interview	10%
Legacy Status	15%

derivation of the strength vector will be explained shortly, it would be valuable to point out that the aggregate strength of an individual candidate's application, ϕ_k, must fall between the following boundary conditions: $0 \leq \phi_k \leq 1$.

Say that $r \in \vec{rating}$. For this case, an individual is assigned unique r values for each of the 7 elements within our simplified model. For a "perfectly perfect" applicant, $s_{rating} = 1$. The opposite is true for a "perfectly horrible" candidate.

Applicant Pool Matrix. Let $M_{pool} \in R^{\Omega \times 7}$ be the matrix representing the weight vector of every candidate in the applicant pool. We initialize all elements of the matrix, $m_{ij} \in M_{pool}$ to have a null value, as we will soon fill them via the "strength" process alluded to earlier.

Race Vector. Lastly, we create another array called $\vec{race}$, indexed to M_{pool}, that enables us to separately track the race of each applicant. This vector will become especially helpful in our *Results* section as we show racial disparities between the status quo for a current Harvard class alongside its predictive post-Affirmative Action equivalent.

B. Derivation (II)

This section will go into great depth to explain the stochastic processes used to determine attribute scores. Naturally, we will follow the same sequence of determination as used in our model.

From our set number of applicants, we establish an iterative loop where we will determine the characteristics of each applicant following an identical process as outlined in the beginning of *The Model* section. The very first thing we determine is the applicant's income.

Income. To determine income in a stochastic manner, we define a helper function, coined $income\_generator()$ that outputs one of $[0, 1, 2, 3, 4]$ corresponding to the bottom 20% of income, 20%-40%, 40%-60%, 60%-80%, 80%+ tiers, respectively. These 5 quintiles are determined by the College Board as follows:

Quintile	$ Range	% of Test Takers
I	<$51,591	12.3%
II	$51,592-$67,083	14.3%
III	$67,084-$83,766	16.3%
IV	$83,767-$110,244	22.2%
V	>$110,245	34.9%

We then invoke the random library, as we will do frequently over the remainder of this process, in order to randomly determine the applicant's quintile (based on the weights seen in the table above). Once determined, the helper function returns the corresponding quintile $[0, 1, 2, 3, 4]$, and we continue.

Race. With our income set, we now must determine the applicant's race. To do so, we define a helper function called $race\_generator(income)$ that takes income as a parameter which influences the eventual output: the applicant's assigned race.

We define six races to divide applicants into (as they are defined in the US Census): White, Black, American Indian, Asian, Pacific Islander, and Hispanic. Mathematical consolidation of certain groups was necessary for the simplification of the model. Such consolidation included redistributing the 2+ race group into the existing categories in an even, weighted fashion, as well as grouping South-Asian into the broad Asian group.

Next, we went about determining each quintile's breakdown by race; that is, each race will have a proportion of the quintile, and these proportions will sum to one. We backed out this data in a roundabout manner given the data sets available to us. We had the proportion of each quintile for each race; for example, the Proportion of Quintile I that was part of the white population. We then multiplied these percentages by the consensus number of this race's total population to find the total number of of each race in each quintile. We could then correctly calculate the weighted proportion of each race in each quintile, which yielded the following table.

Quintile:	I	II	III	IV	V
White	50.78%	56.04%	66.92%	70.72%	71.64%
Black	17.02%	15.15%	11.86%	8.11%	8.22%
Am Ind	1.71%	01.40%	1.13%	0.89%	0.37%
Asian	5.22%	4.74%	06.26%	08.83%	10.32%
Pac Isl	0.36%	0.34%	0.26%	0.18%	0.22%
Hispanic	24.90	22.33%	13.57%	11.27%	9.13%

From there, we established the list of the proportion of each race in Harvard's most recent applicant pool, with the following corresponding values: [White: .406, Black: .152, American Indian: .029, Asian: .279, Pacific Islander: 0.008, Hispanic: .126]. We pulled these directly from Harvard's most recently admitted class, with the necessary redistribution as discussed earlier.

We now have our two necessary tools to stochastically assign race: quintile race distribution and Harvard Application race weighting. Next, we take the aggregate of these two tools to yield our necessary quintile and Harvard Application adjusted race probability weights. We do this by evaluating the following equation for each of the race's proportion within the quintile:

$$\text{Breakdown} = 5\left[\frac{\%\text{ race in given quintile}}{\Sigma\%\text{ of race in quintiles}}\right](\%\text{ of race in Harvard Pool})$$

This equation may not immediately seem intuitive, so let's discuss. First, we calculate the total proportion of a certain

race across all quintiles (which we will use in the next step). Then, we see that the term in the brackets gives us the relative weight of the quintile for the race (i.e. % of white people in Quintile I / total % of white people across all quintiles). We then standardize this across all five quintiles by multiplying this by five. Then, we reconcile the difference between the Harvard Applicant Pool and our data by scaling our existing standardized quintile data by the Harvard data, by multiplying them together. The end yield a 2D matrix with the same shape as our original quintile matrix, but now reconciled with the correct Harvard Application weights.

From here, the hard work is largely over: we query to the corresponding row of the newly created matrix that represents the income quintile that we fed into the $race\_generator(income)$ function at the start. Then, using these queried values, we probabilistically determine the race of the applicant using the same random function as when determining income, but with the new probabilities that we worked to derive.

Standardized Test Scoring. Now, we have the applicants income, which determined it's race. As mentioned prior, race will now sit independently for the rest of the model, until it's time to analyze our results. Let's use our income to help derive our stochastically determined SAT score.

To get our SAT score, we call a helper function, $score\_generator(income)$, which takes in income as a parameter and outputs an SAT score (on the 1600 scale). Taking a dive into our helper function, we define a list of average SAT scores by quintile as: [I: 914, II: 965, III: 1007, IV: 1059, V: 1161]. These mean scores by quintile are as reported by the College Board.[5]

A problem has presented itself: these are the national average SAT scores, and Harvard's SAT scores are significantly higher. Our best adjustment to such is to take a "Harvard Scaling Factor," determined to be 1.5, such that the median quintile mean score (1007) multiplied by said scaling factor yields the median SAT score at Harvard (1510, per Harvard's Admissions). We then establish a list of average SAT scores by quintile for Harvard Applicants (who have a realistic chance at admission), by scaling all items of the previous average list by our "Harvard Scaling Factor." This yields [I: 1371, II: 1448, III: 1510, IV: 1589, V: 1600].[6]

Now, we add a variance factor into these average scores by capitalizing on the normal distribution of SAT scores. We calculate the standard deviation of Harvard Applicant SAT Scores (not national average scores) to be 70, based on reported percentile scores (provided by Harvard Admissions). From here, we calculate a deviation away from the mean of the applicant's pool's quintile using another helper function: $deviation(std\_dev)$.

The $deviation(std\_dev)$ function uses basic laws of the normal distribution to stochastically assign a deviation away from the mean based off of its parameter of the standard deviation. First, we use the random function to see whether we are deviating above or below the mean (at $50/50$ odds). Then, we make use of the $68\% - 95\% - 99.7\%$ rule, which states that for a normally distributed function, 68% of the data lands within one standard deviation of the mean, 95% are within two standard deviations, and 99.7% are within three standard deviations. To implement this, we call on the random function again to see if we fall within one, within two, or within three standard deviations. We will ignore anything outside of three standard deviations for simplicity. Once our standard deviation range has been determined, we then calculate our location within that standard deviation range (i.e. whether we are 1.8σ when within two standard deviations but outside of one or actually only 1.2σ). We then return this value (multiplied by 1 or -1, depending on whether we are above or below the mean) back into our $score\_generator(income)$ function.

We have our mean Harvard Applicant SAT score for our income quintile, and we have our deviation $(+/-)$ from the mean. We sum these two to get out SAT score, and return it.

Now that we have our actual SAT score, we standardized our SAT rating on a [0, .25, .5, .75, 1] scale. The breakdown of score rating is as seen in the table.

These benchmark scores to determine standardized weight were determined by looking at percentile SAT scores of admitted students. We now have our first characteristic that determines admittance: Standardized Test Score.

Std Rating	$ SAT Score
0	< 1350
.25	$1350 - 1430$
.5	$1430 - 1500$
.75	$1500 - 1560$
1	> 1560

Academic Performance. There is a very strong correlation between SAT Score and academic performance. We will take advantage of this when quantifying and standardizing applicant academic performance.

However, we do recognize that there certainly exists variation between SAT Score and GPA (think of your friend that studied for hours but couldn't nail the SAT, or your friend who you never saw pick up a book who thought the SAT was a breeze).

To account for these (and more normal, less extreme variations), we add a deviation factor. To do this, we take use of another random function, $random.uniform(a, b)$, which returns a random number on $[a, b]$ (with all numbers being equally likely to be picked). We called $random.uniform(-.1, .1)$, meaning that our deviation was anywhere in that range.

To calculate our academic rating, we then reference that candidate's SAT rating (remember, these took the values on [0, .25, .5, .75, 1]) and added the random deviation. We corrected to make sure that any academic score greater than one was set back to 1, the maximum. We have now calculated the second characteristic that influences admittance: Academic Performance.

[5] *SAT Suite of Assessments Annual Report.* College Board, 2022, https://reports.collegeboard.org/sat-suite-program-results

[6] We set a cap at a SAT score of 1600 for quintile V here.

Before we move on to our next characteristic, let's take a quick view-back on what we have determined so far: (See *Figure 1*)

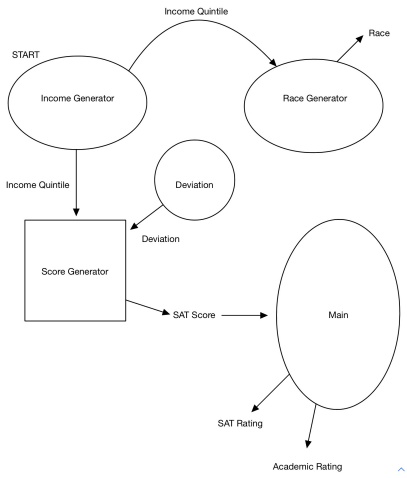

Fig. 1. Stochastic System Current State Flowchart

Geographic Location. Geographic location certainly has some influence over who gets into Harvard, but there are far too many factors to be able to dive deep in-depth and relate geography to other academic indicators. Therefore, we simply assign a standard rate of anywhere from [0, .25, .5, .75, 1] with an equal probability of each.[7]

Extracurricular Activities. Extracurricular Activities have long been the crux of the college admissions process for your classic student. However, prestigious universities (Harvard certainly included) have begun to place more weight onto applicant's participation in such activities.

Our methodology for determining the score of a student's extracurriculars is not over complicated, but there are some significant changes compared to that of prior characteristics. Firstly, we do not tie extracurriculars to SAT Score or Academic Performance. We believe that the correlation between the characteristics, though positive, is low enough to omit and treat as independent variables. We follow the same general format score it on the standard [0, .25, .5, .75, 1] scale based off stochastic factors. However, we have altered our probabilities to more accurately reflect the nature of extracurricular participation distribution among applicants.

As per the nature of Harvard, many applicants sit near the very bottom (perfect academics, with little to no other involvement) and the very top of this rating (student body presidents, leaders of several clubs and volunteer organizations). Further, we believe that most of the applicants to Harvard are generally motivated within their school communities, so approximately half of the applicants will be assigned an above average rating. Therefore, our probabilities of being assigned to the score ratings are as follows:

Std Rating	$ Probability
0	30%
.25	10%
.5	10%
.75	30%
1	20%

We then evaluate the standard rating of extracurricular activities for the applicant as we have done so prior, and we now have four of our determining characteristics.

Personal Essays. The personal essay rating strategy is very akin to that of rating extracurriculars with some key differences.

We will follow the same general probabilistic weighting process as before: we assign ratings based on determined probabilities. However, instead of using the standard rating scale, we are going to use a scale of [0, .25, .5, .75, 1, 2]. The inclusion of a 2 as the highest rating is a reflection of the power of an extremely emotional personal essay that resonates with your reader.

Further, we believe that applicants have started to realize the importance of crafting a near-perfect essay to submit as a part of their application (as evident by the increase of demand in the market for essay review tutors). As a result of this, the overall quality of these essays will be strong, with fewer essays being rated as truly poor, and a high percentage of essays being rates in the very good (but not quite great) range. Considering these two probabilistic factors, our probability distribution is as follows:

Std Rating	$ Probability
0	15%
.25	15%
.5	30%
.75	30%
1	6%
2	4%

With the applicant's essay rating being stochastically selected based off of these probabilities, we now have $\frac{5}{7}$ of our deterministic characteristics.

The Interview. Though an interview is far from the most important part of a candidate's application, it can prove a pivotal role in making, or breaking, their admission chances. The college interview has a lot of variation, though it normally breaks down into only four results: Your interview did not like you, they felt indifferent towards you, they did like you, or they *really* liked you. As a result of this, we will assign rating to those circumstances as: [-.5, 0, .5, 1]. The reason that the gap in these ratings is larger is due to the polarization of the interviewer: if they sort of don't like you or only kind of like you, they will likely reflect both of these with a near neutral rating.

Another key factor to interviews is that not every applicant is guaranteed an interview. We will assume that anybody who

[7]This is perhaps one of the only true limitations of the model, but we felt that it would be nearly impossible to quantify the relation that geography had to all other characteristics.

does not receive an interview will score a rating of true neutral. To account for this, we simply add more neutral ratings of 0 into the ratings scale; instead of our original rating scale, we will use: [-.5, 0, 0, 0, .5, .1].

From here, we will assume an equal probability of obtaining any of the indices within the newly altered rating scale (that is, you have a $\frac{3}{6}$ probability of receiving a neutral 0 rating on your interview). After stochastically assigning your interview score, we now have just one more characteristic to consider.

Legacy Status. Harvard is known for leaning their bias towards applicants with a legacy status. Meaning, if an applicant's mother, father, or close family relative went to Harvard, they have a better chance of getting in.
The legacy status is perhaps the most unique in terms of our stochastic process, as there really only seems to be options: you either are a legacy or you're not. However, we consider another case: a "super-legacy," which we consider to be any applicant who's a legacy whose parent has donated a very large sum of money to the school. Once past a certain threshold, donations by a parent can *significantly* increase their legacy child's chance at admission. So much so, that we assign the following values as legacy scores [No Legacy: 0, Legacy: .75, Super-Legacy: 3].

Now, a legacy is fairly rare: we are assuming that 90% of applicants are not legacies, with 9.98% being normal legacies, and a mere 0.02% of applicants being "super legacies." We then assign the legacy score rating to an applicant using the same stochastic process that we have used throughout the rest of the model. And finally, we are done defining an individual applicant's characteristics.

C. Derivation (III)

Before pushing onward, we need to define several additional data structures which will enable us to interpret our findings.

Strength Vector. The strength vector, $\vec{strength} \in R^{\Omega \times 1}$ is a horizontal flattening of M_{pool} to one-dimension. By summing every element within a fixed row, for every row, we can determine the relative "strength" of every applicant's admissions profile. As alluded to earlier in the derivation section, the collection of all Ω strength numbers, denoted as ϕ_k (where $k \in Z$ between 0 and Ω), comprises $\vec{strength}$.

Note that the next four expressions are all vectors initialized with null values. Their length is determined as a consequence stochastic process; thus, it is directly dependent upon the probabilistic nature of the model itself. The significance of these vectors will be explained below.

Admitted Vector. Let α represent the number of admitted students on Ivy Day, the date on which Harvard releases admissions decisions. The admitted vector, $\vec{admitted}_{all} \in R^{1 \times \alpha}$, holds the strength number of every accepted candidate whose $\phi_k > c.$, where c represents a cutoff value that will soon be defined.

Acceptances by Race. The acceptances by race vector, $\vec{admitted}_r$ also lives in $R^{1 \times \alpha}$. However like $\vec{race}$, its values are indexed to the race of each admitted pre-frosh. Therefore, entries into $\vec{accept}_r$ will *key* to one of the six previously defined racial groups, as outlined in the U.S. Census.

Committed Vector. The committed vector, $\vec{committed}_{all}$, represents all ϕ_k values of the incoming class. Our admissions model accounts for students who refuse their offers of admission, which is what distinguishes $\vec{committed}_r$ from $\vec{admitted}_{all}$. Not everyone who gets in decides to come. Although we did not implement a functional wait-list, which could be an interesting addendum to the model, we were able to stochastically predict the number of students who would accept vs. deny their offers of admission. This will be explained in further depth later-on.

Commitment by Race. Like its $\vec{admitted}_r$ counterpart, the commitment by race vector, $\vec{committed}_r$, stores the race index of every person listed in $\vec{committed}_{all}$. If β people refuse their offers of admission, then $\vec{committed}_r$ and $\vec{committed}_{all}$ both live within $R^{1 \times (\alpha - \beta)}$.
Now, we can begin to discuss the predictor function, which is the bread and butter of our model. The lone input of the function is a cutoff value, denoted as c, which determines the amount of admits in an admissions cycle. Essentially, $0 < c < 1$, so the cutoff value lives in the same domain as each ϕ_k. While iterating through every ϕ_k in $\vec{strength}$, the predictor function determines which students achieve admission in a rather binary fashion.

1) If $\phi_k > c$, k-th student gets in.
2) If $\phi_k \leq c$, k-th student is denied.

The beauty of the strength vector and cutoff value approach is its quantitative nature. If a student's strength vector is *good enough*, they are accepted within a matter of microseconds. Else, they are rejected within the time frame and can move on with their college search.

By creating four bins, initialized at 0, to keep track of admitted students, denied students, committed students, and dropping students, respectively, we can begin to log our admissions results.[8]

After a student has been accepted, the admitted students counter is incremented by one, $\vec{admitted}_{all}$ is appended with the k-th student's ϕ_k, and $\vec{admitted}_r$ is appended with the key to that same student's race.

In the event that a student was denied, the denied student's counter would be incremented by 1. That said, we are not particularly interested with who gets denied; the bulk of the analysis will be placed on who actually achieves admission.

[8]"Dropping students" conglomerates the total value of denied students with the number of students who turned down their offer of admission.

In our tertiary step from the initial round of admissions statistics, we can calculate the rough acceptance rate for an incoming class, contingent upon c.

Once accomplished, we insert another probabilistic condition into our model, which predicts whether a student will elect to attend Harvard, given that they've already gotten in. Let's call this the faux yield rate, p_w, which in our model was held constant at 0.9.

For every one of the α admits, we generate a random number, τ between 0 and 1 where all possible fractions are uniformly possible. If $\tau > 1-p_w$, then the k-th student accepts their offer of admission. Else, they deny it.

If the offer is accepted, the committed student counter gets incremented by 1, and their ϕ_k is added as an element to $\vec{admitted}_{all}$, as their race key is to $\vec{admitted}_r$. If a student rejects their offer, the dropped students counter increases by 1; the counter includes both the initial denials and post-acceptance rejections. Finally, by finding the ratio of committed students to accepted students, we can determine the true yield rate amongst accepted students.

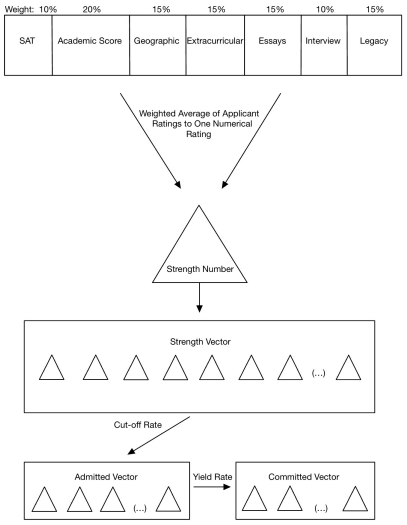

Fig. 2. Consolidation of Strength into Committed Class

This concludes the formal derivation of our primary-phase admissions model. In our concluding section, we will touch upon possible improvements and manipulations that could be made to our system, but for now, we have provided a theoretical schematic through we can compare our own world to the hypothetical.

Next, we will attempt to answer the questions outlined in our introduction, one-by-one.

III. Results

With the potential impact of Students for Fair Admissions vs Harvard looming on the horizon, we are pleased to report that with the aid of our theoretical model, we have answered all of our paper's initial questions. We share our findings below:

What value of c generates a class size most comparable to that of the incoming class of 2027 at Harvard?

Over the last four years, Harvard's admitted class sizes have been $1,980$ (2024), $1,968$ (2025), $1,954$ (2026), and $1,942$ (2027), respectively. Taking these data points into account, we found the average class size to be exactly 1,961 students. //

In order to determine what cutoff value would yield the closest admitted class size to Harvard's recent average, we generated a plot displaying the dependence of class size upon cutoff rate (*See Figure 3*): [9]

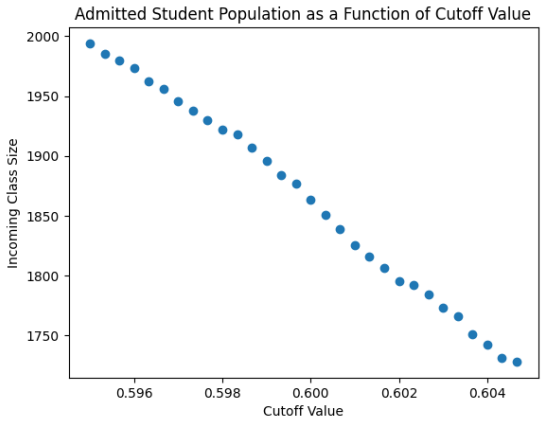

Fig. 3. Impact of Cutoff Value on Admitted Class Size

We have concluded that a cutoff rate between 0.595 and 0.597 produces the most accurate result, and subsequently, the closest acceptance rate to Harvard's actual one.

What are the respective yield and acceptance rates for this value of c?

Taking $c = 0.595$ and running our simulation a a multitude of times, we determined the average yield and acceptance rates for various strength vectors to be roughly:

1) Acceptance Rate: $3.21\% \pm 0.2$
2) Yield Rate: $90\% \pm 0.7$

Our theoretical yield of 90% was selected with our lack of wait-list dynamics in consideration. We were unclear as to how many students would be transferred from the wait-list to the accepted pile in a given admissions season. This data is of course publicly available, but it varies a great deal year-to-year and is very difficult to predict. As a result of this, we decided to increase our theoretical yield, from .85 to .9 (with variation), to make up for no wait-list admittances.

[9] https://college.harvard.edu/admissions/admissions-statistics

What does the frequency distribution of ϕ_k-s look like in an admitted and committed student class?

Despite the somewhat convoluted composition of the elements of $M_{pool} \in R^{\Omega \times 7}$ and its simplified form, $\vec{strength} \in R^{\Omega \times 1}$, the data in all three graphs appears to follow a Gaussian distribution. We have deduced that this is because Central Limit Theorem, which is a concept that we expound upon in our conclusion.

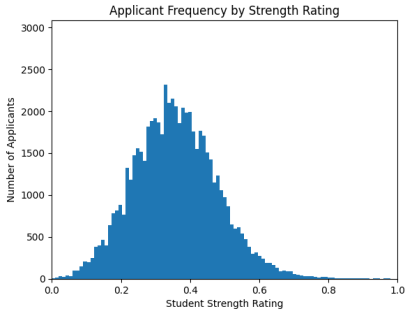

Fig. 4. Applicants vs. Strength Rating

The bell curve centered at a value below 0.3 might initially seem surprising, however, it is a direct consequence of our decision to weight our various application categories distinctively. If all categories had been amplified equally, we could have had more confidence in producing a mean nearer or precisely at $\phi_k = 0.5$.

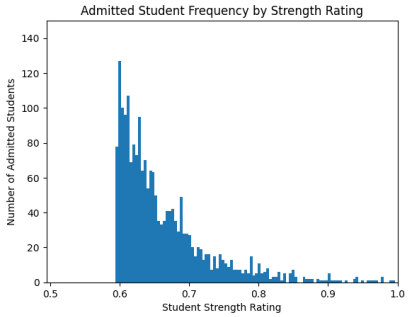

Fig. 5. Admitted Students vs. Strength Rating

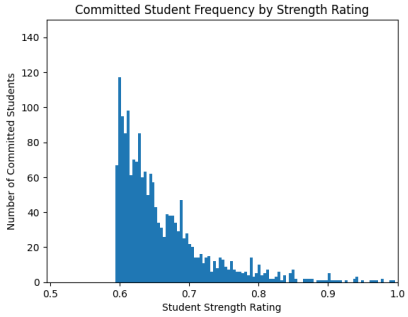

Fig. 6. Committed Students vs. Strength Rating

At the cutoff value, we see all applicants with a strength number south of $\phi_k > c$. vanish.

In addition, it is also worthwhile to identify that adding every bin tied to every nonzero c should yield the total number of students in the model's accepted class.

After the students who rejected their offers of admission are removed from the dataset, we see a similiarly curved figure with slightly sparser bin values.

IV. Conclusion

Our model of the college admissions process independent of affirmative action turned out to be an overwhelming success. We were able to predicate an approximately normal distribution of candidate strength ratings (as per the Central Limit Theorem, which states that the average of independent variables tends to be normally distributed). Although not all of our variables can be considered independent and that our average has been weighted, we can still evidence of this idea (See *Figure 4, Figure 5, and Figure 6*). From this, we were able to provide a fundamental range of our cut-off threshold as to where we would decide to accept candidates to yield a class size within the range of a standard committed Harvard class today (≈ 1950).

Our largest question, and the primary motivator for this paper, still remains:

Which racial group(s) will be hurt most by the repeal of Affirmative Action? Which (if any) will benefit as compared to the current admissions system?

Let's first take a look at the admitted students by race as compared to the committed students by race (see *Figure 7* and *Figure 8*, respectively).

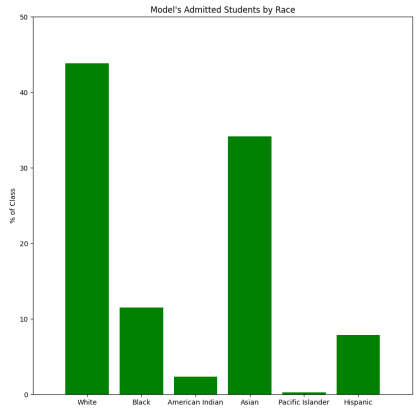

Fig. 7. Model Class Accepted Students by Race

The distributions of accepted students vs race and committed students vs race appear to be nearly identical. This is attributable to the implementation of the yield rate from accepted to committed student being independent of

any of the applicant's characteristics. In real life, this may not be so simple: students with exceptionally strong ratings may have admission offers from equally prestigious schools. However, since the yield rate is so high, this discrepancy is nearly negligible. Now, let's compare our models to actual data from the Harvard Class of 2024 (see *Figure 9*).

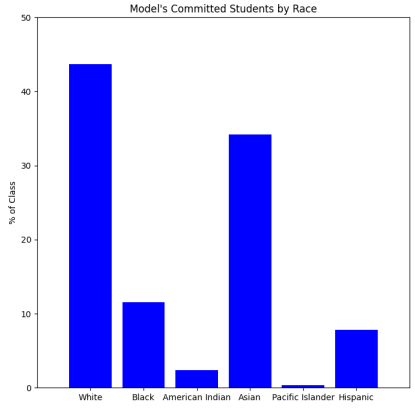

Fig. 8. Model Class Committed Students by Race

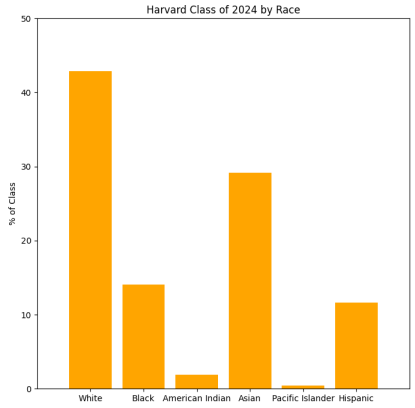

Fig. 9. Harvard Class of 2024 by Race

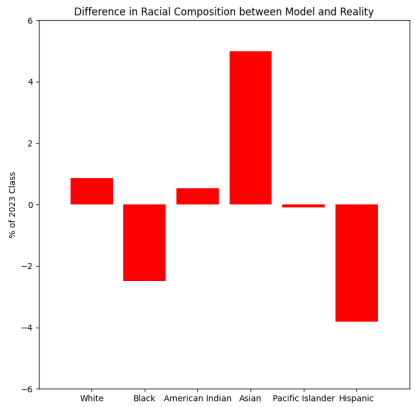

Fig. 10. Difference in Racial Composition between Model and Harvard

By observation alone, we see that the proportion of white students increases, the proportion of Black students decreases, the proportion of American Indian students increases, the percentage of Asian students increases significantly, the percentage of Pacific Islanders increases slightly, and Hispanic proportion takes a big hit. Quantifying these changes in class proportion: [White: +.9%, Black: -2.7%, American Indian: +.5, Asian: +5.1%, Pacific Islander: +.1%, Hispanic: -3.9%].

Based off of this quantitative data, it appears that the in a post-Affirmative Action system, the groups that would take **the hardest hit in representation would be the Hispanic and Black populations. Pacific Islander representation remains essentially flat. Both American Indian and White representation are up, but by far the biggest benefit to any group goes to the Asian population**, up nearly 5%.

The purpose of this paper, as stated in the introduction, was to create a hypothetical college admissions system that could function in a post-Affirmative Action reality, not to decide whether Affirmative Action was correct or not. That said, we can deduce the value of Affirmative Action (and the effects of removing it) from our model. Affirmative Action, by intention, increases representation for historically under-represented groups.[10] Yet in doing so, colleges end up also under-representing other populations, specifically people from Asian backgrounds. The fight over what is fair has placed Affirmative Action practices underneath the magnifying glass and in front of the Supreme Court.

If we had more time to expand the model, we would look to explore the following in greater depth: Test-optional functionality, wait-list dynamics, and a more rigorous geographic ranking distribution.

For test-optional functionality, we realize that this is becoming increasingly prevalent in the admissions process given the current disdain for standardized testing catalyzed by COVID. We reflected this counter-momentum in our model by only assigning a 10% weight to test scores. However, we acknowledge that this is a "broad fix" approach, and if we were to build up from square-zero again we would add this as an option. Of course, our bottoms-up approach starts with our test score, as we derive test scores stochastically from income. Then, we scale these scores to derive GPA (with a deviation). Thus, test-optionality would require us to change the roots of our model, which might prove a difficult fix.

We touched upon our lack of wait-list dynamics in the results, but as a summary: wait-list removal variation can be very high year-year. To combat this, we increased our theoretical yield, from 0.85 to 0.9 (with variation), to compensate for no wait-list admittances.

Finally, for the geographic ranking distribution, we did attempt to modulate this based off of SAT score, but noticed that the correlation was not very strong. Beyond this, linking geographic influence across the other characteristics would

[10]These discrepancies are evident in the race-income quintile breakdowns.

have caused a multitude of headaches, ranging from certain locations having a better chance of receiving an interview (alumni density) to legacies being far more concentrated in certain areas than others throughout the country.

Despite our model's limitations, we hope that with a process of revision, our simulated system might prove a useful tool in the future evolution of college applications systems, regardless of whether race is taken into consideration or not. In the meantime, we, alongside the rest of the United States, will anxiously await the outcome of Students for Fair Admissions v. Harvard, which is expected to be released in the coming months.

ACKNOWLEDGMENT

The authors would like to thank Professor Cengiz Pehlevan and the rest of the AM50 course staff: Shanshan, Emin, and Matthew. We feel that our mathematical modeling abilities have grown by leaps and bounds this semester, and we are grateful for the time and effort invested into our class. Thanks again, and have a wonderful summer!

REFERENCES

[1] For access to our team notebook, which holds all of the code for our model, please click on the following link: https://colab.research.google.com/drive/1Rxy8fySrUxIIgtU43vxm0g BPTE$_2 l9gW?usp = sharing$

[2] *Digest of Education Statistics, 2021. National Center for Education Statistics (NCES) Home Page*, a Part of the U.S. Department of Education, nces.ed.gov/programs/digest/d21/tables/dt21$_2$26.50.$asp?current = yes$. Accessed 8 May 2023.

[3] *SAT Suite of Assessments Annual Report.* College Board, 2022, https://reports.collegeboard.org/sat-suite-program-results

[4] Hlr. *"Students for Fair Admissions, Inc. v. President amp; Fellows of Harvard College."* Harvard Law Review, 24 Mar. 2023, harvardlawreview.org/print/vol-134/students-for-fair-admissions-inc-v-president-and-fellows-of-harvard-college/.

[5] Mears, Bill. *"Selected Quotes from Supreme Court Affirmative Action Arguments."* CNN, 9 June 2013, www.cnn.com/2013/06/07/politics/court-affirmative-quotes/index.html.